DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Graph-based Support Vector Machines for Patient Response Prediction Using Pathway and Gene Expression Data

Citation

Huang, Norman Jason. 2013. Graph-based Support Vector Machines for Patient Response Prediction Using Pathway and Gene Expression Data. Doctoral dissertation, Harvard University.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:11169763

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Graph-based Support Vector Machines for Patient Response Prediction Using Pathway and Gene Expression Data

A dissertation presented

by

Norman Jason Huang

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

June 2013

©2013 Norman Jason Huang
All rights reserved.

Dissertation Advisor: Professor Cheng Li Norman Jason Huang

Graph-based Support Vector Machines for Patient Response Prediction Using Pathway and Gene Expression Data

Abstract

Over the past decade, multiple function genomic datasets studying chromosomal aberrations and their downstream implications on gene expression have accumulated across a variety of cancer types. With the majority being paired copy number/gene expression profiles originating from the same patient groups, this time frame has also induced a wealth of integrative attempts in hope that the concurrent analysis between both genomic structures will result in optimized downstream results. Borrowing the concept, this dissertation presents a novel contribution to the development of statistical methodology for integrating copy number and gene expression data for purposes of predicting treatment response in multiple myeloma patients.

This dissertation is structured in three complimentary sections. The first reviews the methods currently available for integrative purposes between gene expression and copy number data. Specifically this includes the conceptual evolution of these workflows, approaches used amongst varying methods, endpoints targeted for downstream analysis, and biological milestones achieved through such efforts. The focus here is to highlight the accomplishments and potential areas for improvement. A key takeaway message is the lack of integrative attempts in the field of response prediction.

The second section consequently introduces a new integrative approach for response prediction. This section is furthermore split into two subsections where the first describes the motivation, intuition, theoretical developments, and simulation/application results with respect to the proposal; while the second describes an extension to include copy number data. Note that since the approach introduced in the initial subsection only utilizes the gene expression data, it will therefore require the latter argument to complete its integrative design.

The final section then concludes the dissertation by discussing future steps in data integration and how these innovations can potentially lead to more efficient and robust response prediction models.

Contents

List of Figures

List of Tables

Acknowledgments

I will like to express my deepest appreciation to my advisor, Cheng, along with my committee members, Tianxi and Dr. Munshi, for their ungoing support and help through this entire process.

Gene Expression And Copy Number Integration

Chapter 1

This chapter reviews methodologies designed to integrate gene expression and copy number data. The discussion entails the evolutionary development of these methods, the different approaches used for integrative purposes, the targeted end-points of downstream analysis, and various biological milestones/novel insights achieved by these workflows.

1.1 Introduction

Human cancer genesis and progression are enabled by the aberrant function of regulatory genes that control aspects of cell proliferation, apoptosis, genome stability, angiogenesis, invasion, and metastasis[1]. This dogma, well established even before the advent of functional genomics, confers the crucial idea that recurrent genomic abnormalities promote an underlying selection advantage by spanning across genes vital for tumor development and metastasis[2]. Amongst these abnormalities somatic copy number alteration (CNA) of oncogenes/tumor suppressor genes (TSG) and their downstream implications on gene expression (dosage effect) have become key events pioneering the discovery of many important biological results.

In particular the concept of dosage effect has been heavily used to identify regulatory genes located within regions of focal or chromosomal level amplifications/deletions. Most notably amplified oncogenes include ERBB2[3], MYC[4], CCND1[5], CAD[6, 7], BCR-ABL[8], and AR[9], while deleted TSGs include PTEN[10], CDKN2A[11], RB1, BRCA1, BRCA2, PTPRJ, and TP53[12–15]. In addition research showing the consistency of CNAs in cancer (average of 24 gains and 18 losses per tumor sample across 26 cancer types) has highlighted the importance of these events[2] thereby making their discovery and functional assessment an essential process to elucidate cancer biology.

To assess these genomic changes and their downstream implications, the past decade has witnessed a dramatic increase both CNA and gene expression (GE) based studies. In addition to the biological insights that have accompanied their arrival, the influx of these new datasets[1] has also captivated all researchers and analysts alike. In particular the ones that contain both sources of genomic information in reference to the same samples have garnered the most optimism and interest of all.

Specifically these 'paired datasets' are valued due to the anticipation behind 'data integration', or its subsequent analysis. From a biological and analytical point of view, this workflow highlights a combined analysis between CN and GE profiles such that it would: (1) Benefit any analysis due to its ability to assess more recurrent aberrations

[1]Most of these datasets are generated using high throughput microarray technology. From the perspective of GE, the maturation process of these experiments has standardized the resulting output whereas numerous options still exist to gather CNA information. For example, both array comparative genomic hybridization (aCGH) and single-nucleotide polymorphism (SNP) microarrays can been used to obtain high-resolution information on CNAs[16, 17]. While SNP-arrays can also be used to detect allele-specific information (chromosomal loss of heterozygosity and unipareitnal disomy), aCGH represents a more traditional karyotyping method reserved for CN purposes. The key, however, is that both methods offer an alternative to conventional cytogenetic approaches in the study of cancer related CNAs[18].

and their corresponding dosage alterations[2]; (2) Increase the accuracy to differentiate between driver and passenger alterations; and (3) Allow for optimal power and a reduction in false positives [19, 20]. With all these aforementioned advantages, it should come as no surprise that they have accumulated an impressive degree of popularity and attention[2].

Table 1: Condensed List Of Integrative Datasets

Cancer Type	Year	GE Information	CN Information	Samples	GEO Accession
Kidney	2009	70; HG-U133A	159; Mapping250K_Sty	229	GSE14994
Multiple Myeloma	2009	158; HG-U133A	45; Mapping50K_Xba240	203	GSE16122
Lymphoma	2008	203; HG-U133_Plus_2	203; Custom	406	GSE11318
Liver	2008	91; HG-U133_Plus_2	197; Mapping250K_Sty	288	GSE9829
Leukemia	2008	81; HG-U133_Plus_2	79; Mapping50K_Hind240; Mapping50K_Xba240	160	GSE10792
Sarcoma	2010	158; HG-U133A	415; Mapping250K_Sty	573	GSE21124
Breast	2010	359; SWEGENE H_v2.1.1 55K	359; SWEGENE_BAC_32K_Full; SWEGENE_BAC_33K_Full	718	GSE22133
Ovarian	2010	68; HuGene-1_0-st	72; GenomeWideSNP_6	140	GSE19539
Lung	2011	100; HG-U133_Plus_2	101; Mapping250K_Nsp	201	GSE28582
Lung	2011	49; Custom	271; Custom	320	GSE31800
Oral	2011	79; HuEx-1_0-st (exon)	122; GenomeWideSNP_6	201	GSE25104
Mesotheliomas	2011	53; HG-U133A	53; Agilent-014693 Human Genome CGH Microarray 244A	131	GSE29211
Breast	2011	197; HG-U133_Plus_2	173; Agilent-014693 Human Genome CGH Microarray 244A	370	GSE23720
Multiple Myeloma	2011	304; HG-U133_Plus_2	254; Agilent-014693 Human Genome CGH Microarray 244A	558	GSE26863
Multiple Myeloma	2010	258; HG-U133_Plus_2	233; Mapping250K_Nsp; Mapping250K_Sty	491	GSE21349

Paired datasets with CN and GE information, similar to the ones depicted here, are commonly found on public repositories such as the Gene Expression Omnibus (GEO). However the paired samples (ones with both CN and GE information from the same patient) are usually only available on a fraction of the study population due to the difficulty associated with their collection process. Note that in both the 'GE Information' and 'CN Information' columns, the first number represents the number of available samples.

With that being said the remainder of this chapter would be reserved for the discussion of integrative methods. Since this concept would eventually see application for response prediction, it lays a robust foundation for the subsequent chapters presented in this thesis.

[2]This upside in integrating CNA and GE profiles is also reflected in large databases (The Cancer Genome Atlas Project[21] and Gene Expression Omnibus[22]) as the storage and production of paired datasets has dramatically increased during this past decade. A small selection of these datasets are provided in Table 1.

1.2 Quantitative Relationship Between CNA And GE In Different Cancer Types

With the apparent upside of these paired datasets firmly ingrained, their analysis falls into the hands of 'integrative techniques'. These methodologies take advantage of the biological links between CN and GE profiles in order to strengthen the downstream analysis. Therefore to understand the methodological and theoretical developments of these workflows, it starts with rationalizing this binding relationship.

As a simple and efficient summarization of the aforementioned link, numerous studies have attempted to quantify CN and GE similarity through the use of correlations within stratified 'blocks' of altered regions. For example Pollack et al.[23] stratified the CN data entries into five blocks: deletion, no change, and low-, medium-, high-amplifications while Hyman et al.[24] stratified into two blocks: amplified and non-amplified. These arbitrary yet intuitive cutoffs have yielded statistically significant block-wise correlations between the CN and GE data[23–30].

Specifically these studies have reported transcriptional changes for 10-63% of genes in amplified regions and 14-62% in deleted regions across a multitude of cancer types. Furthermore they have also shown that a relative gain/loss in genomic content would increase/decrease the averaged expression levels across all genes in the implicated regions[25, 28, 29]. In breast cancer for example, a 2-fold change in copy number (CN) was linked to a 1.5-fold change in the averaged GE levels[23]. Ultimately the impact of CNAs on the averaged GE levels can be described as a widespread phomena despite the fact that many genes within these altered regions are unrelated to the malignant progression of the cancer.

In the context of individual genes however, these correlation trends will often cease to exist. For example in regions of large gains, significantly downregulated genes can still be commonly found. This was particularly true in prostate cancer as 14% of downregulated genes appeared within regions of amplification while 9% of upregulated genes appeared in regions of deletion[25]. Furthermore even amongst chromosome arms amplified in its entirety, there could still exist contiguous regions where genes are expressed at normal levels[25]. Since numerous regulatory mechanisms, in addition to CNAs can all affect mRNA transcription, these caveats were expected to some degree. Nevertheless they serve as a constant reminder of the limitations in CN and GE integration. Ultimately the analysis driven procedure will only expose part of a complex biological picture.

1.3 Classification Scheme For Integration Methods

Due to the sheer number of integrative methods developed over the past decade (see Table 2 for an example), the critical analysis of these workflows starts with the classification of these seemingly unique efforts. For example upon initial inspection, methods can be grouped into three distinct classes based on the interaction between their biological and methodological complexity. First, there exists a group of 'stepwise' methods (discussion provided in Section 1.4) that typically employ relatively simple techniques for the computation. Not surprisingly their endpoints will also be more intuitive - i.e. trying to quantify the interaction between CN and GE on a global scale. Later developments of these 'stepwise' methods will then take advantage of these previous formulations to target better defined endpoints - i.e. clustering and gene searching. On the other hand there also exist a class of 'joint' methods that are computationally more involved (discussion provided in Section 1.5). Although some of these methods still reference routine biological endpoints, others can be more ambitious - i.e. survival and response prediction. A summarization of this initial classification scheme can be seen in Figure 1.

Table 2: Condensed List Of Integrative Methods

Methodology	Type	Endpoints	Statistical Tools Used
Ace-it	S	Gene Targets (Dosage Effect)	nPHT
Berger et al.	J	Gene Targets (Dosage Effect)	SVD; Gene Shaving
Magellan	S	Exploratory Analysis; Clustering	ES; nPHT; CA; GO
SIGMA2	S; J	Exploratory Analysis; Gene Targets (Correlation)	ES; CA; PHT
SODEGIR	S	Gene Targets (Correlation)	Own Statistic; nPHT
Schafer et al.	S	Gene Targets (Dosage Effect)	CA; nPHT
iCLUSTER	J	Clustering	Latent Variable Model; VS
Van Wieringen et al.	S; J	Gene Targets (CNA Induced DEG)	Own Statistic; BF; nPHT
Akavia et al.	J	Gene Targets (Drivers)	BF; Networking
remMap	J	Gene Targets (Correlation)	RA; VS
DR-Integrator	J	Gene Targets (Correlation)	CA; PHT
CNAmet	S	Gene Targets (Correlation)	Own Statistic; nPHT

A condensed list of available methods for CNA and GE integration. **Integration type**: S (stepwise); J (joint). **Statistical tools used**: ES (exploratory statistics), PHT (parametric hypothesis test), nPHT (non-parametric hypothesis test), CA (correlation analysis), RA (regression analysis), GO (gene ontology), VS (variable selection), BF (bayesian framework); SVD (singular value decomposition).

In addition to the previous categorization scheme, integrative methods can also be grouped in terms of their objective and structure (though somewhat similar to classification based on complexity). Generally speaking this second classification scheme is based on the combination of integrative approach ('stepwise logic' or 'joint logic'), and biological objective (gene/gene-set discovery, subtype classification, etc...).

In the context of these biological endpoints, gene/gene-set discovery methods aim to identify candidate genes[19, 31], pathways[19, 32, 33], and regulators involved in tumorigenesis[34–37]. Thus they attempt to shed light on tumor biology through the identification of prognostic and/or therapeutic related targets[38–41]. Subtype classification methods on the other hand are usually designed to identify patient subgroups exhibiting similar genomic patterns. Thus they attempt to improve disease course prediction by identifying groups with similar prognostic and response to treatment properties[42–45]. A figure summarizing this second classification scheme can be seen in Figure 3 of Section 1.5.

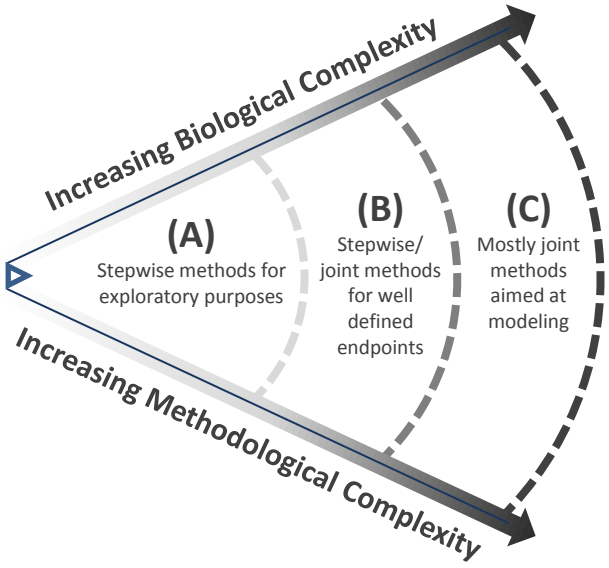


Figure 1: Complexity Of Methodology VS. Complexity In Biological Findings

Most integrative methods can be categorized into three distinct classes based on their biological and methodological complexity. **(A)** Initial stepwise methods designed to explore the relationship between CN and GE employ relatively simple techniques to quantify this interaction on a global scale. **(B)** Later methods will take advantage of this established relationship to search for important genes/genesets or conduct clustering analysis. **(C)** Finally, there is also a class of joint methods that are mathematically involved. Though some may still have well defined biological endpoints, others can be more ambitious (i.e. perturbed pathways). It is important to note however that this classification schema isn't absolute.

1.4 Simple Integration Schemes: Stepwise Integration Methods

The aforementioned 'stepwise integration' approach refers to a class of integrative techniques that structure their analysis according to a biologically sound blueprint. These methods are typically defined by transforming an accepted biological statement into a two-step procedure that together forms the analysis plan. Example statements include:

'CNAs (step 1) can result in differential GE (step 2)', or, 'concordant amplification (step 1) and overexpression (step 2) are tell-tale signs of oncogenes while deletion (step1) along with underexpression (step2) indicate TSGs'. In both statements, the corresponding integrative method will first identify aberrant chromosome regions (step 1) before combining results from a separate expression analysis (step 2) to arrive at the desired endpoints[24, 28, 30, 35, 38, 46, 47].

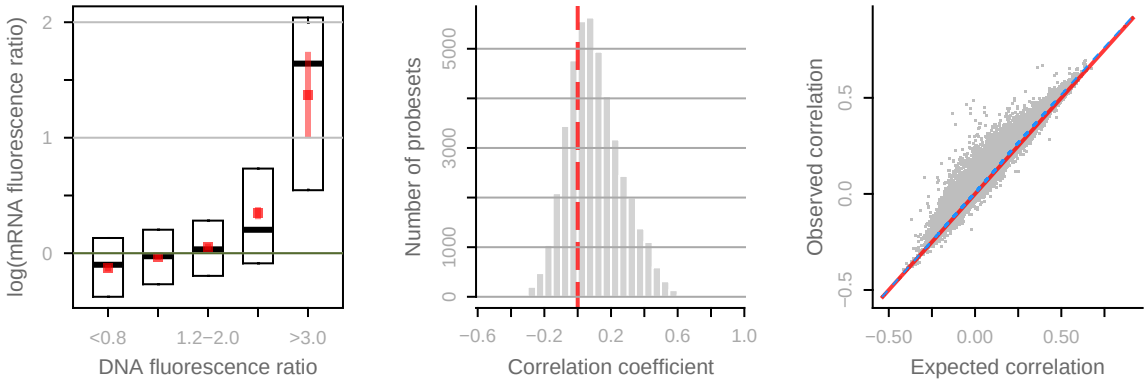


Figure 2: Exploratory Integrative Example

The workflow introduced by Pollack et al. was applied to IFM-I dataset (introduced in Section 5.2). As a result, the genome-wide effect of CNAs on expression levels was explored in multiple myeloma. Since the code corresponding to the original method was not provided, manual programming was required to replicate the workflow. The results were then compared with the original paper to confirm the validity of the replication.

(Left Panel): Boxes indicate the 25th, 50th, and 75th percentile of mean-centered mRNA fluorescence ratios (on $\log_2$ scale) for five classes of probesets stratified accordingly. The five classes were heavy deletion (< 0.8), no change (0.8-1.2), low amplification (1.2-2.0), medium amplification (2.0-3.0), and high amplification (>3.0). The red dot and bar indicates the mean value and confidence interval (95%) within each group. (Medium Panel): Distribution of correlations between CNA and GE profiles for all probesets genes across 282 MM samples. (Right Panel): Plot of observed versus expected correlation coefficients. The expected values were obtained by randomly perturbing the sample labels in the CN data set. The line of unity is indicated. In all three panels, the changes in CNA have a large, pervasive, and direct effect on global expression patterns.

From an application point of view, these stepwise methods are generally used for exploratory purposes (i.e. quantifying CNA and GE similarities on a global scale) and gene identification problems. In reference to the later endpoint, these genes are in reference to the targets implicated in tumorgenesis process. For example ACE-it[48], a stepwise method, identifies genes with concordant CNA/GE relationship. To do so the method initially stratifies all samples into two groups based on CN gain or loss. Afterwards a one-sided Wilcoxen test will assess the degree of concordance with the GE data. Schafer et al.[49] also designed a gene identification method for drivers behind disease progression. In their setup, externally centered correlations are used to assess the degree of concordance between CN and GE

alterations before conclusions are drawn on the gene's driver potential. Other similar methods include workflows by Garraway et al.[50], Wolf et al.[28], and the Stepwise Linkage Analysis of Microarray (SLAMS) algorithm proposed by Adler et al.[33] Through a similar deduction technique involving initial CN classification and later differential expression analysis, they uncovered novel cancer biomarkers and potential regulator genes respectively.

As seen in the setup of these previous examples, stepwise integrative workflows were produced by simply combining the analysis techniques corresponding to the CN and GE data. Not surprisingly the simplicity associated with such design popularized this through process resulting in an abundance of stepwise based methods. However despite their popularity, these techniques are prone to drawbacks. First, since integrative methods require matched CN and GE data at the gene level, added filtering, imputing, and averaging of features were required to account for probeset-loci and resolution differences. While these steps have become conventional necessities for an analysis, they nevertheless add noise to an already complex picture. Second, since many stepwise methods simplify their analysis by introducing arbitrary stratification thresholds to the CN and GE data (call categories or calls), the optimality of the analysis is consistently compromised. Especially in cases where simple calls were used in downstream analysis, results should be taken with extra caution due to the call's inability to account for cancer-related heterogeneity[51].

With the discussion of these stepwise based drawbacks, it is clear that the complexity associated with cancer cells and their corresponding genomic representations will require specialized attention. Consequently oversimplification should be avoided despite the findings that have been uncovered by these workflows; because in the long run, this tradeoff between biological reality and computation feasibility would eventually catch up and undermine the quality of endpoints that could otherwise be extracted.

1.5 Advanced Integration Schemes: Joint Integration Methods

Unlike stepwise methods where the analysis is split into two complementary parts, joint integration approaches are defined by one encompassing workflow. This single analysis highlights the paired relationship between CN and GE entries such that the two genomic sources are treated as a coherent dataset instead of separate structures that each require their own attention.

Since joint integration methods draw conclusions based on the signals that will only emerge as a result of combining both levels of data, it should not come as a surprise that they consistently take advantage of computational tech-

niques allowing for dual inputs. For example numerous groups have implemented correlation based approaches[19, 26, 52–54] and/or regression analysis[27, 55, 56] for a wide variety of biological endpoints. These could range from simple exploratory figures to more complex models targeting response and survival prediction.

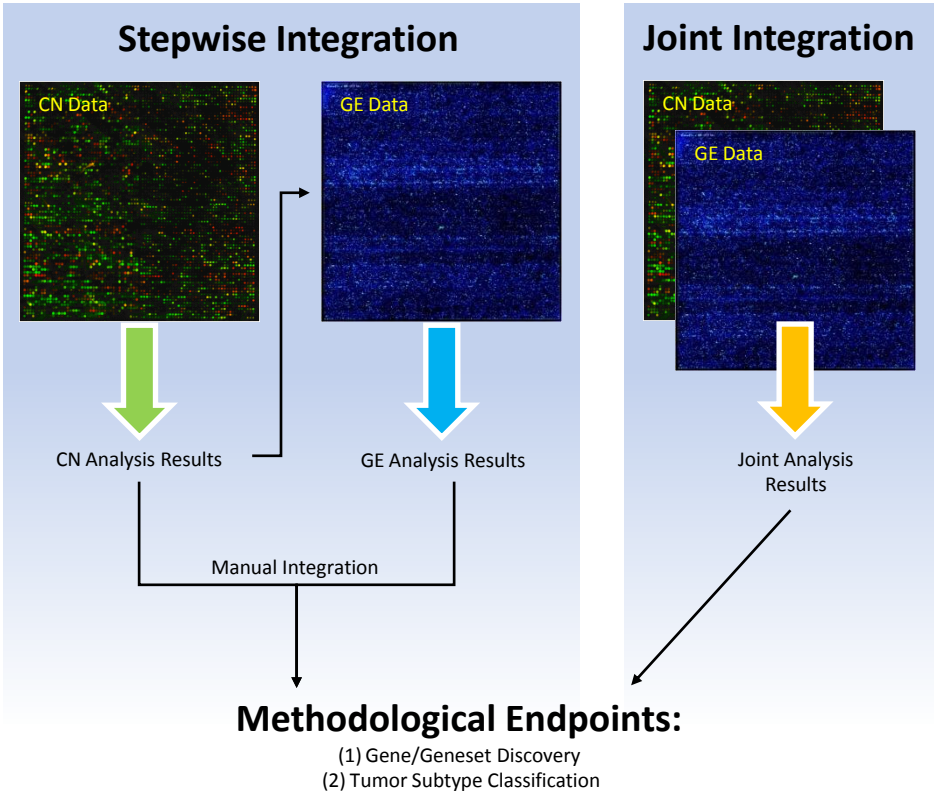


Figure 3: Schematic Overview Of Methods

Integrative methodologies can be grouped based on their integration structure and biological endpoints. Stepwise methods typically interrogate the CN data for regions of CNAs before results from a subsequent GE analysis are manually combined to complete the integrative procedure. Joint integration treats CN and GE as paired data entries. Thus only one analysis is carried out in light of the pairing. Despite the contrasting approaches, most integrative methodologies arrive at the same biological endpoints of gene/geneset discovery or tumor subtype classification.

As a result of emphasizing the binding relationship between CN and GE data, joint methods are generally more comprehensive when compared to stepwise methods. While this improvement is desired from a biological point of view, it however comes at a cost. Specifically, joint methods will often face challenges associated with high dimensionality and computation feasibility due to the need to simultaneously model both data types. For example the imbalance between sample size and feature (number of genes), a problem native to all genomic analyses, will be exacerbated under a joint setting since the added data type essentially doubles the number of features without increasing the sample

size. Thus without a proper treatment of the infused high dimensionality, the burden of so many features can outweigh the benefit of integrative analysis and lead to faulty inference.

To therefore deal with the high dimensionality, joint methods will typically resort to heavily regularization or data reduction. For example Generalized Singular Value Decomposition (GSVD) is popular amongst joint methods due to its added value of dimension reduction. This strategy was implemented by Berger et al.[57] to identify variation patterns between CNA and GE inputs by iteratively projecting both data types onto different decomposition directions. Additional details of their method can be seen in Table 4.

The use of data reduction can also be seen in correlation-based applications. In general correlations are a main staple for joint integration since they effectively modeled two data types (CN and GE in this case) simultaneously. However due to the aforementioned issues of high dimensionality, they can not be directly applied without modification. For example Soneson et al.[58] employed Principal Components Analysis (PCA) to first achieve dimension reduction before Canonical Correlations (CC) identified a set of highly correlation genes. Similarly Gonzalez et al.[59] implemented regularized CC to explore the binding relationship between CN and GE. Their use of regularization was also meant to target the high dimensionality of the analysis. Other methods implementing a similar strategy include: The Significant Overlap of Differentially Expressed and Genomic Imbalanced Regions algorithm (SODEGIR)[60] and the workflow introduced by Schafer et al.[49]. Additional details can be seen in Table 4.

An extension of these early correlation based methods was the concept introduced by Lee et al.[19] for purposes of capturing 'distant' correlations. Normally joint methods based on correlations only emphasize the relationships at the gene level. In other words, the computation is restricted to paired the CN and GE entries corresponding to each gene. However since multiple genes can be coexpressed throughout the genome and similarly, multiple CNAs can simultaneously occur at different locations[38], the previous limitation will restrict the ability to capture all interacting CN/GE relationships - or essentially the distant relationships. Therefore to model this feature, Lee et al. proposed a correlation analysis by allowing clusters of coexpressed genes to be associated with CNAs through the genome. By implementing a bi-clustering algorithm on the observed CN/GE correlation matrix, the workflow was therefore capable of linking genes with significant correlations to the CNAs of other genes. In doing so, all significant relationships between the CN and GE data regardless of their positioning in the genome could be identified.

While improvements such as the aforementioned ones prompted the growth of correlation based joint techniques, there are still drawbacks in light of these efforts. First many of the correlation methods, designed to assess linear trends (i.e. Pearson correlation), are unfortunately not the most practical or robust measures of dependency. For example

since detectable correlations required data points to exhibit a certain degree of spread, the presence of clustering even in the extreme regions will expose the inability of these methods to pick out the corresponding features. Furthermore since the necessary linearity between the CN and GE profiles is hardly a guarantee, the relevance of these methods can also be questioned. Second, since correlation methods model the complex within- and between-data relationships, or essentially the gene-to-gene and CNA-to-GE correlation matrices, they are almost guaranteed to employ judicious assumptions as a means to curb the high dimensionality involved with the modeling process. As a result the risk of oversimplification, unreasonable inference, and faulty results can all be potential byproducts. And finally, despite the ability of correlation based methods to target a wide variety of biological problems, their application is still limited to unsupervised learning endpoints. Though not entirely a surprise, correlations resemble exploratory measures and are therefore less suited for analyses structured by predefined targets and objectives.

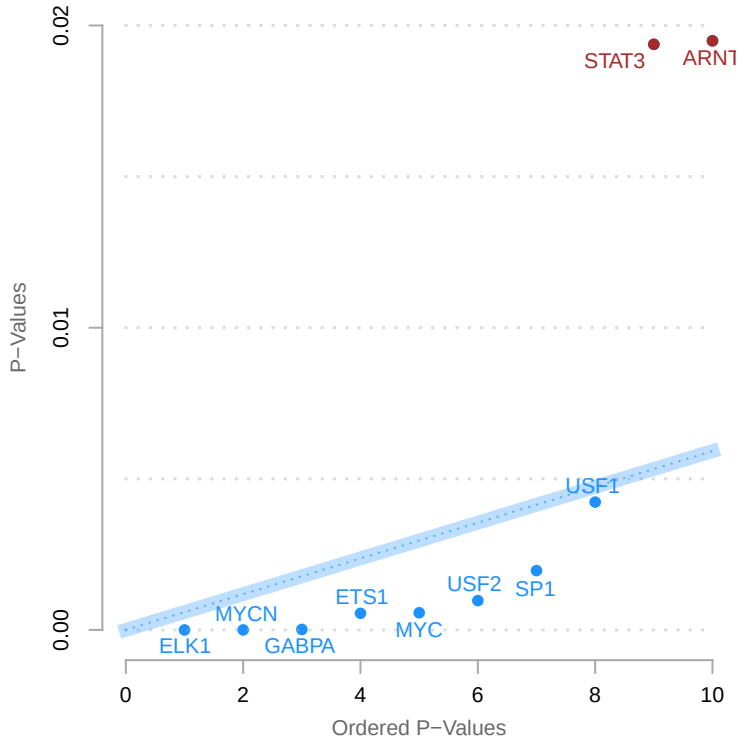


Figure 4: Correlation Integrative Example

CN and GE samples from 73 liver cancer patients (GSE11318) were analyzed using DR-Integrator. Correlation methods similar to the one implemented here would typically identify a list of significantly correlated genes that could be used for downstream functional analysis. In this case the returned genes were studied for a significant presence of TF targets. After adjusting for multiple hypothesis testing (FDR as indicated by the blue boundary), the significantly enriched TFs were plotted and indicated in as blue points (with blue labels). Note that of the 169 TFs analyzed, only 8 turned out to be significantly enriched amongst the list of significantly correlated genes.

To deal with supervised learning endpoints, joint integrative methods have typically turn to regression based techniques. For example Shen et al.[42] introduced a latent variable regression approach for tumor subtype classification/prediction. By modeling the subtypes as latent variables, inference was conducted by simultaneously capturing genomic patterns that were: Strong and consistent across both data types (CNA and GE); Strong but specific to either data type; Weak yet consistent across both data types. In particular the last genomic pattern targeted signals that will only emerge as a result of integrative analysis[42]. Additional details of their method are provided in Table 4.

At the bottom line joint methods emphasize the representation of all genomic inputs as one complementary picture. And while this exemplifies a more comprehensive biological approach, it also adds strain to the computation and modeling process (in comparison to stepwise methods). Consequently this raises questions regarding their purpose - are they truly intended to discover novel biology or just another mathematical exercise? Because on one hand joint methods despite their complexity, offer more reality while on the other, stepwise methods, while favoring intuition, also run the risk of oversimplify an intricate biological system. This trade off can be seen in Figure 3.

1.6 Important Biological Findings From Integrative Analysis

In the past decade numerous biological novelties have come to light through the integrative work between CN and GE profiles. Amongst these findings the majority of the success can be attributed to the identification of genes and pathways altered during tumorigenesis. Specifically some of the notable findings for gene-related endpoints include:

- WHSC1L1 and TPX2 amplification in pancreatic ductal adenocarcinoma and nonsmall-cell lung cancer (Tonon et al.[61]);
- MITF as a potential 'lineage addiction' oncogene necessary for cancer development and progression in a variety of tissue types (Garraway et al.[50]);
- RUNX3 deletion in breast cancer (Chen et al.[62]);
- VEGFA overexpression resulting from 6p21 amplification in heptocellular carcinomas - the pioneering work in oncogene activation via. noncell-autonomous mechanisms (Chiang et al.[44]);
- CSN5 and MYC as genetic regulators in breast cancer (Adler et al.[33]);
- NCOA2 as a nuclear receptor coactivating oncogene in prostate cancer (Taylor et al.[63]);
- NCSTN and SCRIB as potential drivers for heptocellular carcinoma progression (Woo et al.[37]);

- TBC1D16 and RAB27A as potential drivers of melanoma (Akavia et al.[34]);
- Various genes (734 in total) from the lymphomagenesis, cell cycle, apoptosis, and DNA repair pathways as differentially expressed and amplified/deleted in T-cell prolymphocytic leukemia (Durig et al.[64]).

And for pathway related endpoints:

- The abnormal regulation of protein trafficking contributing to proliferation in melanoma (Akavia et al.[34]);
- The amplification of 7p13 with significant correlation to the expression values of genes within the epidermal growth factor signaling pathway for glioblastoma multiforme; deletion of chr 13q with NF-kB cascades in bladder cancer; and amplification of chr 11p with reck pathway in breast cancer (Lee et al.[19]).

In addition classification related endpoints have also witnessed a great deal of success. They include tumor subtype classification and patient group identification (ones that exhibit similar survival and/or response traits). A selection of these findings include:

- Classification of intestinal and diffuse subtypes of gastric cancer based on the immunopositivity of ERBB2 and MUC1 (Myllykangas et al.[65]);
- Clustering of breast tumors into three subtypes based on (1) cell line differences, (2) concordant amplification and overexpression of HER2/ERBB2 (also associated with poor survival), and (3) amplifications at the end of chr 17q (Shen et al.[42]);
- Clustering of lung tumors into four subtypes based on (1) deletion of chr 8p and underexpression of EGFR and DUSP4, (2) amplification of chr 12q, (3) degree of deletion of chr 8p, and (4) degree of mutation of EGFR (Shen et al.[42]).
- Identification of poor prognostic group in breast cancer patients that exhibited additional resistant to preoperative paclitaxel and 5-fluorouracil-doxorubicin-cyclophosphamide chemotherapy combination (Zhang et al.[41]);
- Identification of poor response group in ovarian carcinomas patients that exhibited amplification of chr 19q12 (containing CCNE1) and chr 20q11.22 to chr 20q13.12 (Etemadmoghadam et al.[40]).

To see the remainder of the biological results, refer to Tables 3 and 4 for the complete discussion. In particular Table 4 details the integrative efforts that focused on methodological development while Table 3 features the experimentally based counterparts.

Table 3: List Of Integrative Methods Based On Experimental Papers

Authors	Type	Statistical Tools Used	Endpoints	Application	Major Findings
Nigro (2005)	J	Clustering; CA	Exploratory; Clustering	Glioblastoma	(1) CHI3L1/YKL-40 associated with chr 10 deletion and poor survival; (2) Chr 10 deletion associated with denome-wide differences in GE
Tonon (2005)	S	ES; PHT	Genes (Dosage Effect)	Lung (NSCLC)	WHSC1L1 and TPX2 are likely amplified targets in pancreatic ductal adenocarcinoma and NSCLC
Bussey (2006)	J	CA; nPHT	Genes (Correlation)	NCI60 Cell Lines	(1) Association between ERBB2 overexpression and 3p CN; (2) Negative correlation between L-asparaginase and CN of genes near asparagine synthetase of ovarian cancer cells
Rinaldi (2006)	S	CA; PHT	Genes (Correlation)	Lymphoma (MCL)	SYK inhibition can be a therapeutic strategy for lymphoma
Tsafrir (2006)	S	ES	Exploratory	Colon	(1) Large regions of CNA correlate with expression levels; (2)~63% of overexpressed genes show gains; (3)~62% of down-regulated genes show loss; (4) Regions with significant dosage effect likely favor tumorigenicity
Yao (2006)	S	ES; PHT	Genes (Amplification)	Breast	Combined aCGH and SAGE analysis can identify novel oncogenes
Chen (2007)	S	PHT	Genes (Dosage Effect)	Breast	RUNX3 overexpression suppressed the invasive potential of MDA-MB-231 breast cancer cells in a matrigel assay
Durig (2007)	S	ES; PHT; nPHT; GO	Genes (Dosage Effect)	Leukemia (T-PLL)	(1) Identified 734 DEGs, some involved in lymphomagenesis, cell cycle regulation, apoptosis, and DNA repair; (2) Downregulation of TSG MTUS1 corresponded with chr 8p deletion
Kim (2007)	S	CA	Genes (Correlation)	Prostate	(1) CN/GE integration identifies genetic alterations in proposed canadidate genes for cancer progression; (2) Metastic samples displayed the most genomic alterations
Kloth (2007)	S	CA; PHT	Exploratory	Cervix	(1) CNAs inferred through SNP and aCGH arrays had 90% concordence; (2) No correlation between CN/GE found at genome-wide level; (3) Significant dosage effect on chr 5p
Olejniczak (2007)	S	ES	Genes (Dosage Effect); Exploratory	Lung (SCLC)	18q21-23 CNV can be used as a predictor for sensitivity of SCLC to Bcl-2 family inhibitors
Chiang (2008)	S	Clustering; PHT; nPHT	Clustering	Liver	Overexpression of VEGFA via. 6p21 overexpression suggests a non cell-autonomous mechanism of oncogene activation
Gallegos-Ruiz (2008)	S	*Ace-it* method	Genes (Dosage Effect)	Lung (NSCLC)	(1) 14q.32.2-33 is a frequently deleted region affecting HSP90 GE significantly; (2) HSP90 downregulation seem to extend survival; (3) HSP90 has clinical impact for NSCLC patients
Myllykangas (2008)	S	Own Statistic; nPHT	Genes (Dosage Effect); Exploratory	Stomach	(1) ERBB2 amplification and gains at 20q13.12, 17q12 can be used to cluster patients; (2) Intestinal and diffuse type gastric cancers have distinct molecular profiles
Scotto (2008)	S	ES; PHT	Genes (Dosage Effect)	Cervix	(1) 20q amplification is an early stage event in CC; (2) HSIL with 20q amplification correlated with persistence/progression to invasive cancer
TCGA (2008)	S	ES; CA	Exploratory	Glioblastoma	~76% of genes within recurrent CNAs have correlating GE
Bea (2009)	S	PHT	Genes (Correlation)	Lymphoma (MCL)	(1) Identified 35 regions of concomitant amplification and overepxression; (2) Revealed list of target genes involved in MCL pathogenesis
Beck (2009)	S	Clustering; PCA; FA	Clustering	Leiomyosarcoma	Identified three distinct LMS subtypes throught integration of CN and GE data
Bergamaschi (2009)	S	nPHT; FA	Genes (Oncogene)	Breast (Basal-Like)	(1) CAMK1D was usually overexpressed when amplified; (2) CAMK1D is a potential amplified oncogene linked to epithelial–mesenchymal transition in breast cancer
Beroukhim (2009)	S	nPHT; FA	Genes (Dosage Effect)	Kidney (ccRCC)	(1) VHL disease associated tumors are more homogeneous; (2) MYC is amplified and overexpressed
Broet (2009)	J	RA; BF	Genes (Correlation; Survival)	Lung (NSCLC)	CN/GE derived signatures are robust predictors of clinical outcome in early stage NSCLC

Table 3 (Continued): List Of Integrative Methods Based On Experimental Papers

Camps (2009)	S	ES; PHT	Exploratory; Genes (Correlation)	Colon	(1) 8q24 amplification leads to MYC and FAM84B overexpression; (2) CNAs resulting from chromosomal aberrations plays a major role in CRC transcriptional degregulation
de Tayrac (2009)	J	CA; PHT; FA	Genes (Correlation)	Glioblastoma	(1) Cis-acting DNA targeted genes may be critical for glioblastoma progression; (2) TSGs PCDH9 and STARD13 may be involved in tumor invasiveness and resistance to etoposide
Etemadmoghadam (2009)	S	ES; CA; nPHT	Genes (Correlation; Clinical)	Ovary	(1) Amplification of chr 19q12 (CCNE1) is associated with poor response to primary trmt; (2) CCNE1 has a cell-cycle independent role in modulating chemoresponse and is a dominant marker for patient outcome
Haverty (2009)	S	PHT; CA; FA	Genes (Correlation)	Ovary	(1) PVTI was amplified and overexpressed; (2)PRKCI and ECT2 are potential drivers; (3)
Oudejans (2009)	S	*Ace-it* method	Genes (Dosage Effect)	Lymphoma (DLBCL)	(1) 18% overexpressed genes are located in amplified regions; (2) 55% of down-regulated genes are in deleted regions
Reid (2009)	S + J	CA; FA	Genes (Correlation)	Colon	(1) 20q gain was strongly associated with TP53 mutation; (2) PLCG1, DBC1, NDGR1 are candidate genes important in colon cancer
Zhang (2009)	S	CA; RA; Clustering; PHT	Genes (Correlation; Clinical)	Breast	Combined CN/GE analysis can refine patient clustering
Astolfi (2010)	S	ES; CA; PHT	Genes (Correlation)	Gastrointestinal Tract (GIST)	(1) 14q23.1 is most frequently deleted and contains potential novel TSGs - DAAM1, RTN1, and DACT1; (2) Mean CN/GE correlation coefficient is 0.115
Green (2010)	S	PHT; nPHT	Genes (Correlation)	Lymphoma (cHL/MLBCL)	9p24 amplification increases gene dosage of PD-1 ligands and their induction by JAK2
GEDI / Hartmann (2010)	S	Own Statistic; nPHT	Genes (Dosage Effect)	Lymphoma (MCL)	(1) Deregulation of proliferation and DNA damage response pathways may be a result of CUL4A, ING1, and MCPH1 deregulation; (2) Deregulation of Hippo pathway may have pathogenetic role in MCL
Hawthorn (2010)	S	Own Statistic; PHT	Genes (Dosage Effect; Drivers)	Breast (IDC)	PPAR-alpha / RXR-alpha are downregulated pathways in tumor samples
Horlings (2010)	S + J	CA; PHT	Genes (Correlation)	Breast	Correlations between gene expression signatures and underlying genomic changes can be used to construct prognosis signatures
Mosca (2010)	S	PHT; nPHT	Genes (DEG)	Leukemia (CLL)	(1) 2 distinct molecular types of 13q14 deletions in CLL may have clinical implications
Northcott (2010)	J	CA; Clustering; nPHT; PCA	Clustering	Medulloblastoma	Identified four subgroups of medulloblastoma with distinct prognosis and molecular profiles
Parris (2010)	S	CA; PHT; RA	Genes (Correlation)	Breast (DBC)	Identified 47 genes and 1 unigene cluster with significant CN/GE correlation
Paugh (2010)	S	ES	Exploratory	Glioma (HGG)	(1) PDGFRA is the predominant target of focal amplification in childhood HGG; (2) PDGFRA amplification and 1q gain are potential initiating events in childhood gliomagenesis
Taylor (2010)	S	CA; PHT; FA	Exploratory; Genes (Correlation)	Prostate	(1) Identified NCOA2 as onfogene in ~11% of tumors; (2) TMPRSS2-EGR is associated with 3p14 deletion
Xu (2010)	J	ES; RA; nPHT	Exploratory; Genes (Correlation)	Oral (OSCC)	(1) Genome CNA accounted for 31% of GE variation; (2) 11q12.2-11q13.3 shows highest CN/GE correlation
Bekhouche (2011)	S	CA; PHT; nPHT	Exploratory; Genes (Correlation)	Breast (IBC)	(1) Genomic profiles of IBCs were as heterogeneous as those of nIBCs; (2) Percent of deleted genes where GE correlate with CN is 7-fold lower in IBCs; (3) Identified 24 candidate IBC-specific genes
Rose (2011)	S	FA; CA; PHT	Exploratory; Genes (Correlation)	Melanoma (SSM/NM)	Analysis questions the linear progression of SSM to NM in melanoma

Continued: A comprehensive list of experimental papers for CNA and GE integration. Refer to the caption corresponding to Table 4 (Page 17) for the keyword guide.

Table 4: List Of Integrative Methods Based On Methodology Papers

Authors	Type	Statistical Tools Used	Endpoints	Application	Major Findings
Phillips (2001)	S	ES; PHT; FA	Exploratory	Prostate	Chr. gain/loss usually results in over/under exp. of genes
Hyman (2002)	S	Own Statistic; ES; nPHT	Exploratory	Breast	CNAs have substantial GE impact
Pollack (2002)	S	CA; PHT	Exploratory	Breast	CN and GE interlocked
SLAMS / Adler (2004)	S	SAM; CA; Clustering	Genes (Regulators)	Breast	Wound resposne signature induced by MYC/CSN5 amplification
Lipson (2004)	J	Own Statistic; CA	Genomic Regions (Correlation)	Breast	Identifies subtypes based on CN/GE correlation
Masayesva (2004)	S	ES	Exploratory	Head and Neck (HNSC)	CNAs alter GE over large chr. regions (some unrelated to cancer progression)
Wolf (2004)	S	ES	Exploratory	Prostate	Overall impact of CN on GE largely attributed to low level gain/loss
Garraway (2005)	S	Clustering; PHT	Genes (Dosage Effect)	Melanoma	Identified MITF as the target of a novel melanoma amp.
Heidenblad (2005)	S	ES	Exploratory; Genes (Dosage Effect)	Pancreas	60% of genes in high amplified regions are overexpressed
Berger (2006)	J	Singular value decomposition; Gene shaving	Genes (Correlation)	Breast	Similar patterns exist between CN and GE data
Chaudhary & Schmidt (2006)	S	ES; nPHT	Exploratory	Prostate	LOH leads to overall downregulation of GE (i.e. tumor supressors).
Chin (2006)	S	ES; nPHT; PHT; SA	Genes (Response); Classification	Breast	Clustering improved by combining CN and GE; Low level CNAs progresses cancer by altering RNA and cell metabolism
Jarvinen (2006)	S	Own Statistic; ES; nPHT; GO	Exploratory	Head and Neck (HNSC)	Amplifications have clear impact on GE
Magellan / Kingsley (2006)	S	ES; nPHT; CA; GO	Exploratory; Clustering; Survival	Ovary	Identifies correlation between GE and clinical outcome.
Phillips (2006)	S	ES	Pathways	Glioma (HGG)	Akt and Notch signaling is poor prognosis for gliomas
Ruano (2006)	S	ES; PHT	Exploratory; Genes (Dosage Effect)	Glioblastoma	Overexpression of amplified genes attributed to gene dosage
Sweet-Cordero (2006)	S	Own Statistic; nPHT; GO	Genes (Drivers)	Lung (Murine)	DNA damage response / telomere maintenance enriched in correlated genes
Ace-it / Van Wieringen (2006)	S	nPHT	Genes (Dosage Effect)	NA	NA
Yao (2006)	S	ES; PHT	Genes (Amplification)	Breast (DCIS)	Combined aCGH and SAGE analysis can identify novel oncogenes
Riccadonna (2007)	S	SVM; Machine Learning	Classification	Breast	NA
Sorocenau (2007)	S	ES; nPHT; PHT	Exploratory	Glioblastoma	Phosphoinositide 3-kinase/Akt pathway important for high-grade gliomas
Stranger (2007)	S	ES	Exploratory	Lymphoblast	CNAs capture 17.7% of GE variation
Yoshimoto (2007)	S	ES; PHT	Exploratory	Kidney (CCC)	CNAs are correlated with GE deregulation in CCCs
SIGMA2 / Chari (2008)	S + J	ES; CA; PHT	Exploratory; Genes (Dosage Effect)	NA	NA
Ding (2008)	S	CA	Genes (Mutation)	Lung	Novel genes mutated
Gu (2008)	J	ES; CA; RA	Exploratory	Breast; Pancreas; Prostate; Lung	Global CN/GE correlation is weak; Segmenting CN improves CN/GE correlation
Lee (2008)	J	CA; Biclustering	Clustering	Bladder; Breast; Glioblastoma	Unique pathways altered in each cancer

Table 4 (Continued): List Of Integrative Methods Based On Methodology Papers

SODEGIR / Bicciato (2009)	S	Own Statistic; nPHT	Genes (Correlation)	Astrocytoma; Kidney	NA
Kotliarov (2009)	J	CA; PHT	Genes (Correlation)	Glioma	Correlation coefficients ranged from -0.6 to 0.7
Menezes (2009)	J	RA	Genes (Correlation)	Breast	Robustness can be improved by studying association between gene sets
Orozco (2009)	S	ES; nPHT; PHT	Exploratory	Metabolic Traits (Murine)	83% of genes within CNVs are DEGs; CNVs may contain regulatory elements for altered GE
Schafer (2009)	S	CA; nPHT	Genes (Dosage Effect)	AML	More robust results in comparsion to standard procedures
iCLUSTER / Shen (2009)	J	Latent Variable Model; VS	Classification	Breast; Lung	Identified novel tumor subclass
Van Wieringen (2009)	S + J	Own Statistic; Own Distribution; BF; nPHT	Genes (Dosage Effect)	Breast	Concrete statistical theory backing up biological findings
Woo (2009)	J	CA; SA; PHT	Genes (Driver)	Liver	Identified 50 genes with specific signaling molecules (mTOR, AMPK, and EGFR)
Akavia (2010)	J	BF; Networking	Genes (Driver)	Melanoma	Abnormal regulation of protein trafficking contributes of melanoma proliferation
pint / Huovilainen (2010)	NA	Own Statistic; CA; BF	Genes (Correlation)	Stomach	NA
remMap / Peng (2010)	J	RA; VS; Dimension Reduction	Genes (Correlation)	Breast	Identify amplified trans-hub region in 17q12-q21 influencing 30 unlinked genes
DR-Integrator / Salari (2010)	J	CA; modified TT	Genes (Correlation)	Breast	NA
Soneson (2010)	J	CA; PCA	Genes (Correlation)	Leukemia	Split sample into well known subclasses using selected genes
CNAmet / Louhimo; Hautaniemi (2011)	S	Own Statistic; nPHT	Genes (Correlation)	Glioblastoma; Ovary	NA

Continued: A comprehensive list of methodology papers for CNA and GE integration. **Type**: S (stepwise); J (joint). **Statistical tools used**: ES (exploratory statistics), PHT (parametric hypothesis test), nPHT (non-parametric hypothesis test), CA (correlation analysis), RA (regression analysis), FA (Functional Analysis), SA(Survival Analysis), GO (gene ontology), VS (variable selection), BF (bayesian framework); SVD (singular value decomposition).

1.7 Guidelines For Using Existing Integrative Analysis Methods

When planning an integrative analysis (between CN and GE data), the appropriate choice of methodology usually plays an integral part behind the quality control of the final results. Thus the dual task of identifying and filtering out methods to obtain the 'best' technique becomes paramount for success. Amongst these two steps the first resembles a shotgun approach where all qualifying methods are noted (an example provided in Tables 3 and 4 of Section 1.6) while the later will then match personal assumptions and dataset characteristics [3] to the most optimal integrative method. It is important to note that the final filtering step isn't limited to the options mentioned here as additional constraints may be included based on personal choice and preference.

Nevertheless the first step, as mentioned, is to generate a working list of methods corresponding to the desired biological endpoint. Though this may seem daunting, it should be pointed out that since all endpoints can be grouped into three families, only three lists theoretically exist. Therefore the initial task shouldn't be overly taxing as it may otherwise seem.

The endpoints can be grouped as follows: Exploratory analysis; Gene identification; and Clustering. First, exploratory methods (i.e. designed to infer the extent of CNA/GE relationship) are highlighted by the workflows of Pollack et al.[23], Hyman et al.[24], Wolf et al.[28], etc... Therefore studies looking for similar biological conclusions can mirror these workflows as methodological guidelines. In terms of the filtering step, since exploratory methods are all relatively similar in construction, their performance shouldn't waiver by much given that a signal is indeed present in the data. Second, there exists another class of methods designed for gene identification. Thus studies targeting individual genes/pathways deemed important due to dosage effect (SIGMA2[66], ACE-it[48]), correlation (DR-I[67], SODEGIR[60], CNAmet[68], remap[69]), and CNA induced differential expression can all resort to these methods. Note that while these examples consistently analyze downstream expression changes, their diverse assumption base and construction process makes filtering nontrivial. Thus unlike the previous category, selecting an optimal gene identification method will require additional effort. Finally, the last class of methods are reserved for clustering purposes. They specifically target endpoints related to the classification of tumor subtypes and clinical groups induced by CNAs and differential GE. iCluster[42] for example, identifies tumor subtypes characterized by concordant CNAs and GE changes while the workflow introduced by Garrayway et al.[50] reaches the same endpoint through the identification of lineage-specific regulators. Again, filtering is required to select the most optimal method corresponding to the

[3] CNA and GE platforms, resolution, preprocessing methods, and final representations.

properties of the dataset and goals set in the analysis.

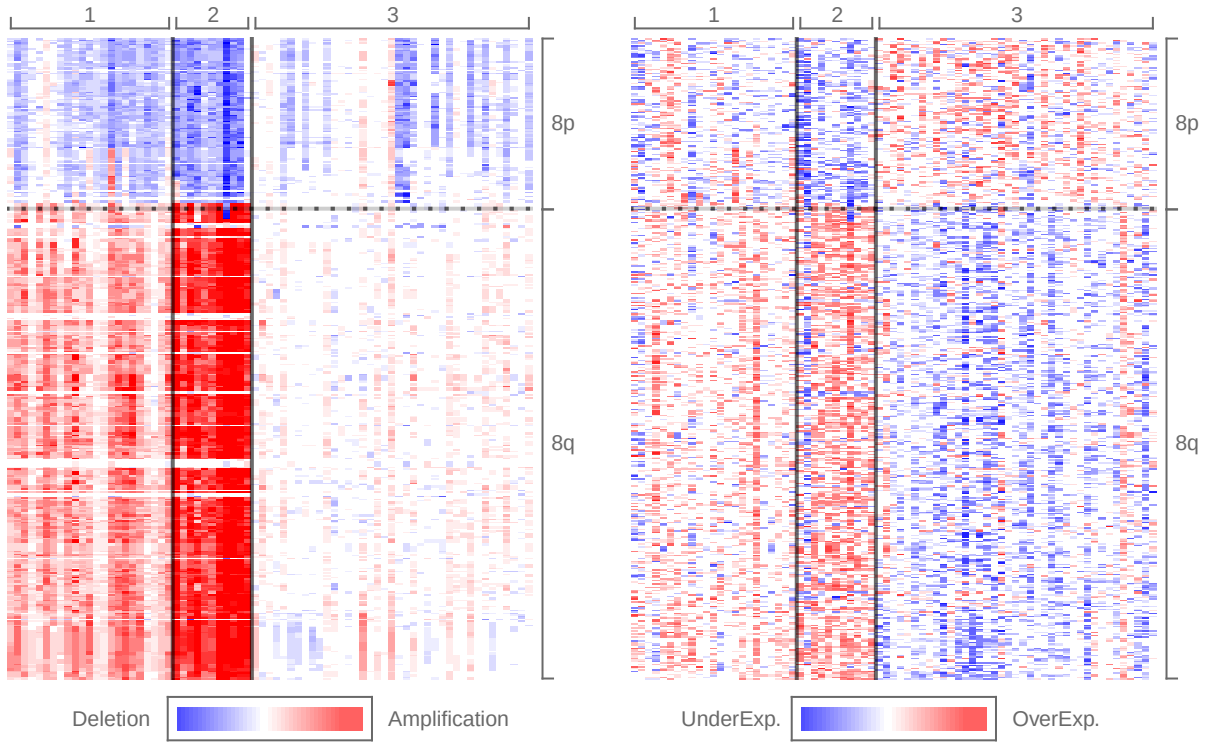


Figure 5: Clustering Integrative Example

73 paired CN and GE samples from liver cancer patients (GSE11318) were classified according to the hidden variable model proposed by Shen et al. The left heatmap indicates the classification results from chromosome 8 imposed on the CN data while the right heatmap indicates the same classification imposed on the GE data. Amongst the three inferred groups, the first was generally classified by light amplification in 8q, the second by concordant deletion/underexpression of 8p and amplification/overexpression in 8q, while the third by minor underexpression of 8q.

Without doubt the selection of an optimal integration method will play a key role for any analysis. While the aforementioned examples only represent a small fraction of potential methodologies corresponding to each endpoint, it should be noted that the brevity here was only meant as a concise summary and should not be used in any application. In practice, additional methods should be noted so that the potential of including the most suitable one is maximized. In regards to the filtering step, since each dataset is unique, filtering is therefore a personalized procedure reserved for the researcher. Consequently without a physical structure laying out its merits and guidelines, it is omitted from this discussion.

1.8 Discussion: Past

Integrative analysis between CN and GE profiles has been become a mainstream analysis approach due to the potential of such workflow. In this past decade the persistence of numerous integrative efforts has helped advance cancer biology and clinical care. After all when $>15\%$ of heritable variation in GE can be attributed to CNAs, it clearly highlights the interlocking nature of both data structures[27]. And in light of such relationship the blueprint for integrative creativity was paved out for all cancer types in general.

Integrative methods have contributed the most to the working knowledge of CNAs and their downstream implications in various cancer types. While incapable of revealing the complete story behind these genomic alternations, their application to defined endpoints has exposed numerous consequences. For example a popular research avenue of identifying the extent of CN and GE relationship has revealed that $\sim 60\%$ of all genes exhibit differential expression concordant to their CN status (suggesting a cis-dosage interaction between these two variables[55]). While dependency summarizations similar to this one are ultimately just ballpark estimates that furthermore vary with cancer type and methodological choice, they nevertheless reinforce the existence of a global correlation trend interlocking both data structures together.

Exploratory analyses similar to the one previously mentioned, are popular due to the widespread implications that usually accompany their results. These sweeping findings, while easy to generate methodologically, have provided deep insight into transcription regulation from a CN point of view. For example the fact that increased expression is a direct response of amplification suggested that most genes are not prone to auto-regulation as a result of dosage compensation. This remains true despite the notion that most genes are incapable of completely overriding the transcription regulatory mechanisms already in place. Another example was the potential role of widespread CNAs in tumorigenesis[70, 71]. Due to the popularity of analyzing data from an individual gene level, alterations that occur on a larger scale are oftentimes taken with less thought and even ignored in many cases. Typical to aneuploidy, these massive gains and/or losses are regularly treated as the byproducts of tumorigenesis from upstream events propelling cancer development - i.e. mutations, CNAs, or regulatory changes to individual targets. However widespread CNAs and concomitant gene expression changes have been shown to disrupt critical stoichiometric relationships in cell metabolism and physiology (i.e. proteosome mitotic spindle). Altogether these factors could promote further chromosomal instability and as a result, contribute to tumor development and progression. From a clinical setting, their impact can also be seen. Since a substantial portion of phenotypic individuality could be traced back to variations in the underlying CNA, their analysis can potentially benefit cancer therapeutics designed to target these imbalances. on a larger scale are oftentimes taken with less thought and even ignored in many cases. Typical to aneuploidy,

these massive gains and/or losses are regularly treated as the byproducts of tumorigenesis from upstream events propelling cancer development - i.e. mutations, CNAs, or regulatory changes to individual targets. However widespread CNAs and concomitant gene expression changes have been shown to disrupt critical stoichiometric relationships in cell metabolism and physiology (i.e. proteosome mitotic spindle). Altogether these factors could promote further chromosomal instability and as a result, contribute to tumor development and progression. From a clinical setting, their impact can also be seen. Since a substantial portion of phenotypic individuality could be traced back to variations in the underlying CNA, their analysis can potentially benefit cancer therapeutics designed to target these imbalances.

In regards to methodological development, integrative techniques have also undergone a gradual transformation over the past decade. From the initial exploratory tools, most methods nowadays have become specialized procedures in hope that the added focus will eventually lead to novel findings and results. In particular the specialization has paid off for purposes of identifying individual genes and patient clusters. For example in the initial category, integration has uncovered numerous targets of CNA, drivers, and subtype-specific genes all vital for tumor formation. For clustering purposes, various analyses have also identified tumor subtypes and patient groups based on the differential clinical characteristics. These may include patient survival and response to therapy measures.

In the near future the development of integrative methods still remains an exciting area of research. As newer data types gather popularity amongst the bioinformatics community, it almost becomes a foregone conclusion that they will eventually play a vital role in the next wave of integrative efforts. Whereas the majority of this current discussion was devoted to published methods, the presented analysis will likewise benefit future integrative attempts as many of the issues that cloud researchers today will most likely still persist as problems in tomorrow's world. For example, (1) The use of efficient dimension reduction will inevitably remain a stressing point due to the high dimensionality involved with any genomic analysis (that is furthermore exacerbated in an integrative setting with added data types); (2) The uncertainty associated with random dichotomizations of the CN data should be propagated in the downstream test statistics or altogether substituted with call probabilities; (3) Prior to the integrative analysis, tumor subclasses should be identified so to reduce tumor heterogeneity; (4) Indirect relationships (interactions between CNA and GE profiles not restricted by physical location) should be accounted for despite the complexity associated with their formulation and modeling process; (5) Gene interaction and regulatory network information should be used similar to a prior on a more consistent basis; (6) Functional enrichment analysis and clinical information should be incorporated directly into the inference process instead of being recognized as a postanalysis interpretation tool; and finally (7) Causal analysis followed by experimental validation should be explored due to the lack of associations that link gene expression and disease directly to a particular CNA.

Together these guidelines and the influx of non-standard genomic data types should pave the way for the next generation of integrative methods.

1.9 Discussion: Future

Without a doubt integrative methods have greatly advanced our understanding of cancer biology from a theoretical and clinical point of view. However despite their accomplishments, the full potential of this workflow has only been partially scratched leaving plenty of room for future developments.

From an analytical point of view for example, there has been surprisingly few cases where CNAs were explored in reference to LOH and UPD. Similarly the role of chromosomal aneuploidy in cancer has also been avoided; though to some extent the question marks surrounding these genomic abnormalities may have contribute to their own stigma. After all when the origins and functions of these events can still be debated[72], their analysis is usually infused with an overwhelming amount of instability. Because if one views aneuplodiy as the central initiator of tumor formation[73–75], then the corresponding analysis would become polar opposites to another treating them as just the side effects of deranged cell division cycles[76, 77]. As a result the positions taken with regards to these questions would largely determine the analysis thus making them overly volatile for attraction. Finally, the area of predicting response to particular therapies (response prediction) has also witnessed few initiatives. Without question the majority of the blame can fall on the complex nature intrinsic to this very problem. Specifically since it has been shown that response is dictated by a variety of subtle mechanisms[78], the ability to do accurate prediction might be a foregone conclusion given that many of the associated data types nowadays are incapable of measuring these delicate changes. To make things even worse, binary classification problems often require sample sizes unattainable in genomic settings. Thus not surprisingly these blunders increase the difficulty associated with this research topic ultimately leaving it with little to no interest.

To answer these analytical questions it is obvious that the starting point involves the development of the next generation of integrative methods. And while a strict interpretation of the guidelines provided in this chapter can guarantee the quality of these new formulations (in terms of efficiency and effectiveness), it will remain doubtful whether or not future workflows can successfully translate into research based models (i.e. cancer aneuploidy or response prediction). After all if the ability to solve a biological question is tangent to the design of the method, then assurance from a methodological point of view becomes a futile attempt to salvage results. For example the statistical power,

noise tolerance levels, and sample size requirements associated with these integrative methods are oftentimes unknown due to the complex nature of genomic datasets. Nevertheless these tangential factors still possess an insurmountable amount of influence on the ability of each method to reach its desired endpoint. Thus a research question (i.e. cancer aneuploidy or response prediction) characterized by unattainable levels to these factors can simply be a lost cause. Because regardless of methodological design, the endpoints are inherently out of reach.

In spite of all these considerations, the future of data integration still maintains a promising outlook. As additional genomic data types gather popularity, new methodological designs will accompany their arrival and eventually lead to the advent of cutting-edge biology. Next generation sequencing, epigenetic methylation, and histone modification for example, are just a few of these notable resources that with maturation, can provide the desired panoramic view of the underlying biology and consequently offer more compelling insights to cancer and its corresponding clinical care.

References

[1] B. Vogelstein, K. Kinzler, *Recherche* **67**, 02 (2002).

[2] D. Albertson, C. Collins, F. McCormick, J. Gray, *et al.*, *Nature genetics* **34**, 369 (2003).

[3] D. Slamon, *et al.*, *Science* **244**, 707 (1989).

[4] K. Alitalo, M. Schwab, C. Lin, H. Varmus, J. Bishop, *Proceedings of the National Academy of Sciences* **80**, 1707 (1983).

[5] P. Hinds, S. Dowdy, E. Eaton, A. Arnold, R. Weinberg, *Proceedings of the National Academy of Sciences* **91**, 709 (1994).

[6] G. Wahl, R. Padgett, G. Stark, *Journal of Biological Chemistry* **254**, 8679 (1979).

[7] R. Schimke, R. Kaufman, F. Alt, R. Kellems, *Science* **202**, 1051 (1978).

[8] P. Koivisto, *et al.*, *Cancer research* **57**, 314 (1997).

[9] M. Gorre, *et al.*, *Science Signalling* **293**, 876 (2001).

[10] J. Li, *et al.*, *science* **275**, 1943 (1997).

[11] I. Orlow, *et al.*, *Journal of the National Cancer Institute* **87**, 1524 (1995).

[12] M. Nagai, *et al.*, *Genes, Chromosomes and Cancer* **11**, 58 (2006).

[13] P. Ra, *et al.*, *Nature* **305**, 779 (1983).

[14] S. Baker, *et al.*, *Cancer research* **50**, 7717 (1990).

[15] C. Ruivenkamp, *et al.*, *Nature genetics* **31**, 295 (2002).

[16] R. Redon, *et al.*, *nature* **444**, 444 (2006).

[17] D. Pinkel, D. Albertson, *Nature genetics* **37**, S11 (2005).

[18] M. Salman, S. Jhanwar, H. Ostrer, *Clinical genetics* **66**, 265 (2004).

[19] H. Lee, S. Kong, P. Park, *Bioinformatics* **24**, 889 (2008).

[20] O. Monni, *et al.*, *Proceedings of the National Academy of Sciences* **98**, 5711 (2001).

[21] N. C. Institute, The cancer genome atlas homepage (2011).

[22] N. C. Institute, Gene expression omnibus homepage (2012).

[23] J. Pollack, *et al.*, *Proceedings of the National Academy of Sciences* **99**, 12963 (2002).

[24] E. Hyman, *et al.*, *Cancer research* **62**, 6240 (2002).

[25] J. Phillips, *et al.*, *Cancer research* **61**, 8143 (2001).

[26] A. Järvinen, *et al.*, *Oncogene* **25**, 6997 (2006).

[27] B. Stranger, *et al.*, *Science* **315**, 848 (2007).

[28] M. Wolf, *et al.*, *Neoplasia (New York, NY)* **6**, 240 (2004).

[29] B. Masayesva, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 8715 (2004).

[30] L. Soroceanu, *et al.*, *Proceedings of the National Academy of Sciences* **104**, 3466 (2007).

[31] L. Qin, *Cancer informatics* **6**, 369 (2008).

[32] H. Phillips, *et al.*, *Cancer cell* **9**, 157 (2006).

[33] A. Adler, *et al.*, *Nature genetics* **38**, 421 (2006).

[34] U. Akavia, *et al.*, *Cell* **143**, 1005 (2010).

[35] A. Sweet-Cordero, *et al.*, *Genes, Chromosomes and Cancer* **45**, 338 (2006).

[36] A. Bergamaschi, *et al.*, *Molecular oncology* **2**, 327 (2008).

[37] H. Woo, *et al.*, *Cancer research* **69**, 4059 (2009).

[38] K. Chin, *et al.*, *Cancer cell* **10**, 529 (2006).

[39] P. Broët, S. Richardson, *Bioinformatics* **22**, 911 (2006).

[40] D. Etemadmoghadam, *et al.*, *Clinical Cancer Research* **15**, 1417 (2009).

[41] Y. Zhang, *et al.*, *Cancer research* **69**, 3795 (2009).

[42] R. Shen, A. Olshen, M. Ladanyi, *Bioinformatics* **25**, 2906 (2009).

[43] C. Kingsley, *et al.*, *Cancer informatics* **2**, 10 (2006).

[44] D. Chiang, *et al.*, *Cancer research* **68**, 6779 (2008).

[45] A. Beck, *et al.*, *Oncogene* **29**, 845 (2009).

[46] Y. Ruano, *et al.*, *Molecular cancer* **5**, 39 (2006).

[47] J. Yao, *et al.*, *Cancer research* **66**, 4065 (2006).

[48] W. Van Wieringen, J. Belien, S. Vosse, E. Achame, B. Ylstra, *Bioinformatics* **22**, 1919 (2006).

[49] M. Schäfer, *et al.*, *Bioinformatics* **25**, 3228 (2009).

[50] L. Garraway, *et al.*, *Nature* **436**, 117 (2005).

[51] L. Merlo, J. Pepper, B. Reid, C. Maley, *Nature Reviews Cancer* **6**, 924 (2006).

[52] Y. Tsukamoto, *et al.*, *The Journal of pathology* **216**, 471 (2008).

[53] D. Lipson, A. Ben-Dor, E. Dehan, Z. Yakhini, *Algorithms in Bioinformatics* **3240**, 135 (2004).

[54] Y. Kotliarov, *et al.*, *Cancer research* **69**, 1596 (2009).

[55] W. Gu, H. Choi, D. Ghosh, *Cancer informatics* **6**, 17 (2008).

[56] R. Menezes, M. Boetzer, M. Sieswerda, G. van Ommen, J. Boer, *BMC bioinformatics* **10**, 203 (2009).

[57] J. Berger, S. Hautaniemi, S. Mitra, J. Astola, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **3**, 2 (2006).

[58] C. Soneson, H. Lilljebjörn, T. Fioretos, M. Fontes, *BMC bioinformatics* **11**, 191 (2010).

[59] I. González, *et al.*, *Journal of Biological Systems* **17**, 173 (2009).

[60] S. Bicciato, *et al.*, *Nucleic acids research* **37**, 5057 (2009).

[61] G. Tonon, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9625 (2005).

[62] W. Chen, *et al.*, *Genes, Chromosomes and Cancer* **46**, 288 (2007).

[63] B. Taylor, *et al.*, *Cancer cell* **18**, 11 (2010).

[64] J. Dürig, *et al.*, *Leukemia* **21**, 2153 (2007).

[65] S. Myllykangas, *et al.*, *International journal of cancer* **123**, 817 (2008).

[66] R. Chari, *et al.*, *BMC bioinformatics* **9**, 422 (2008).

[67] K. Salari, R. Tibshirani, J. Pollack, *Bioinformatics* **26**, 414 (2010).

[68] R. Louhimo, S. Hautaniemi, *Bioinformatics* **27**, 887 (2011).

[69] J. Peng, *et al.*, *The Annals of Applied Statistics* **4**, 53 (2010).

[70] R. Li, *et al.*, *Proceedings of the National Academy of Sciences* **94**, 14506 (1997).

[71] D. Rasnick, P. Duesberg, *Biochemical Journal* **340**, 621 (1999).

[72] J. Marx, *Science* **297**, 544 (2002).

[73] R. Li, A. Sonik, R. Stindl, D. Rasnick, P. Duesberg, *Proceedings of the National Academy of Sciences* **97**, 3236 (2000).

[74] P. Duesberg, *Cancer genetics and cytogenetics* **143**, 89 (2003).

[75] I. Shih, *et al.*, *Cancer research* **61**, 818 (2001).

[76] D. Zimonjic, M. Brooks, N. Popescu, R. Weinberg, W. Hahn, *Cancer research* **61**, 8838 (2001).

[77] H. Lamlum, *et al.*, *Proceedings of the National Academy of Sciences* **97**, 2225 (2000).

[78] L. Van't Veer, R. Bernards, *Nature* **452**, 564 (2008).

Response Prediction Overview

Chapter 2

This section presents an overview of response prediction and the methods required for development for this thesis work. The discussion starts by introducing response prediction and the difficulties associated with this research topic. The bridge with data integration is also provided. Afterwards the guidelines to a new modeling technique featured in this thesis will also be presented. As described, the novelty only considers the GE data as Chapter 6 will then deal with its extension to copy number profiles.

2.1 Introduction And Bridge With Data Integration

Cancer chemotherapy has witnessed a great deal of progress ever since the introduction of nitrogen musters and folic acids in the 1940s. These earliest forms of drugs once administered to all individual patients have gradually evolved into targeted therapies specifically tailored for each cancer type. As a result modern day cancer chemotherapy has become a collection of preoperative treatment strategies administered on a cancer specific basis.

However despite the added dimension of specializing treatment, the response of individual tumors to various drugs is still unfortunately, not uniform. In some patients the biological system is just more capable of adapting to the therapy than in others, even when the tumor histologies are identical. As a result this poses a considerable clinical dilemma because patients exhibiting the resistance quality can be spared exposure to radiation or DNA-damaging drugs and instead, be referred to other treatment options to increase survival chances (i.e. primary surgery or dose-intensified protocols). Thus the clinical challenge of identifying molecular markers predictive of response (or treatment toxicity) becomes paramount in order to reduce the variability associated with current treatment strategies.

The identification of these cancer related markers, i.e. response predictors and signatures, has become an area of research accumulating rapid growth and popularity. Due to the promise of precise, objective, and systematic classification, the past decade has yielded a plethora of these tools ranging from endpoints in prognosis[1–6], survival[7–10], to tumor subtype classification[11–15]. Recently DNA microarray-based gene expression profiling (GEP) has also been introduced to this field due to its ability to simultaneously study the expression activity of multiple genes. Combined with research showing that (1) GEPs of cancer cell lines correlate with drug activity[13, 16, 17]/radiosensitivity[18]; and (2) expression signatures predicting sensitivity to chemotherapeutic drugs in vitro can also be used as accurate clinical markers for these drugs in vivo[19]; it shouldn't come as a surprise that such technological advance has opened up a window to identify the next generation of molecular signatures for therapeutic prediction.

To this day numerous efforts to develop these response markers have accumulated. Most notably they are highlighted by the applications to breast[20], esophageal[21], and colon cancer[22]. In these examples, gene expression signatures indicative of response to pre- and post-operative chemotherapy were identified and subsequently formulated into accurate prediction models. These efforts have reinforced the potential of microarray technology to serve as a base for building prediction tools corresponding to any administered therapy.

With the keyword being 'potential' however, the promise of developing response signatures based entirely on GEP is still a proposition despite all of the optimism and apprise garnered. The cause for this uncertainty can be better un-

derstood from a biological point of view of. Specifically it has been well documented that response to any anticancer agent is often the combined consequence of multiple mechanisms[23]. Since gene expression represents only one of these influential factors, considering it all by itself will significantly diminish the chances of finding a useful signature. Furthermore, in light that some predictive signals are inherently undetectable by gene expression profiling (i.e. subtle mutations that do not cause resulting downstream chances in expression levels[23]), the overly naive idea of using GEP alone becomes even more of a foregone conclusion.

As a result of these aforementioned reasons, GEP based response signatures have yet to deliver on their full potential in spite of the successful applications mentioned earlier. In fact amongst the studies that have explored this topic, the greater majority have noted the inability of expression data to effectively model response[24–26][4]. While the echoed skepticism is indeed troubling, the construct of this thesis tends to think otherwise. From a more optimistic point of view, the lack of concrete ideas in response prediction implies an abundance of room for additional research. As a welcoming attraction, this will offer a greater opportunity to easily explore novelties that may be harder to come by elsewhere.

With that being said, since the initial portion of this thesis was devoted to data integration, a way to explore both ideas simultaneously is to model response from an integrative point of view. In additional to the practical considerations of doing so, the proposal also makes sense for a variety of analytical reasons (i.e. many of the problems facing response prediction can potentially be solved by data integration). First, since data integration has been shown to strengthen downstream analysis[27], the apparent uncertainty associated with GEP based response models can be potentially alleviated upon its implementation. Second, since a multitude of genomic mechanisms can all dictate response[23], integrative analysis will therefore enable a combined input resulting in a more complete predictive platform. Third, as a consequence of the second point, the combined input from various genomic sources will override the need to model each data type independently and thus only require one encompassing analysis. And finally, since relatively few studies have explored response prediction using integrative techniques, the idea will also represent a novel contribution in comparison to the efforts that already exist.

---

[4]Studies conducted by this group, despite applications on the same cancer types and drugs, have claimed that GEP is simply incapable of consistently predicting treatment outcome. To make things even worse, many of the so claimed successful signatures, despite the accuracy attested in the initial tumor study, were never validated on subsequent datasets to confirm the predictive power as claimed. Due to these mixed results, it remains unclear whether such an approach can effectively handle response prediction problems in general.

2.2 Outlining SCIRP

To develop a new response prediction tool predicated on data integration, the inadequacy with current methods was initially analyzed. In doing so these persistent shortcomings can hopefully be avoided and as a result, benefit the diagnostic potential and clinical relevance of the new technique. With that being said, the limitations corresponding to current signatures can be roughly grouped into two categories based on their origin. First, there are deficiencies that surfaced due to inherent properties of the data. And second, there are also pitfalls resulting from the incorrect analytical decisions committed from a methodological standpoint.

From a data perspective, the inability of GEP to consistently predict treatment response has been well documented (in Section 2.1). Since the expression data alone is insufficient for prediction purposes, the idea of incorporating additional sources of genomic information becomes the best solution. However when the majority of response-based studies are limited to expression data and occasionally CN profiles, this solution becomes impractical. Thus the restricted access to additional datatypes hinders the freedom to include any input base. Therefore to benefit all studies equally, added data types will need to be: (1) Complimentary to existing CN or GE profiles and (2) Available without additional work. Only by following these guidelines will the resulting method be highly applicable regardless of the rigid data structures in place.

Amongst the available sources of genomic information, biological pathways available through online databases (KEGG[28], Biocarta[29], Reactome[30], NCI[31], etc...) represented an interesting option. From a practical point of view, they are a free source of validated information readily available for use. While this is more of a methodological requirement, their true advantage is seen biologically. Because based on previous studies, generating expression signatures from pathways that potentially affected how a cell responds to a drug has shown promising results[23]. Thus this indicates that tumor responsiveness probably doens't just depend on the expression levels of one or a few genes. Instead, methods that allow comprehensive interrogation of genetic pathways probably hold greater promise to deliver the desired signatures. As a consequence of the groupwise treatment of genes within pathways, it automatically warrants their inclusion. The only question that remains is how to complement them with the existing GE profiles; all which will be answered in Chapter 3 when the full details of the method are introduced.

Switching to the methodological standpoint, there are also numerous factors that can inhibit the performance of GEP based response signatures. For example methods that required unattainable sample sizes, unreasonable signal

to noise ratios, and questionable preprocessing techniques can all limit the information extracted from the data[5]. Not surprisingly the idea of predicting response based on insufficient information will only exacerbate an already difficult task. Complementary to these requirement-based-drawbacks are the poor decisions that together shape the existing workflow. Specifically incorrect assumptions placed on the data and misguided approaches towards the problem all act like counterfactuals that will only work outside the confines of true biology[6]. While they nevertheless structure an analysis, their erroneous nature also nixes the ability to tease out any useful signatures.

To address these shortcomings, the developed technique in this thesis work will approach response prediction as a series of pieced together solutions. First, in response to the requirement-based-drawbacks, heavy regularization will be used so that all features, regardless of the sample size, can participate in the modeling process. Due to the subsequent benefit of avoiding dimension reduction, the need for preprocessing techniques is therefore omitted. Second, by implementing a structure that accommodates multiple assumptions and approaches towards the problem (i.e. prediction based on means, correlations, and additional frameworks manually specified by the user), a versatile yet robust assumption base was used to handle the underlying biology. In doing so a more realistic modeling approach was adopted. By resolving some of the proposed limitations, the hope is that improvements can be seen in the resulting applications.

With the overall goals of the method in place, the remaining work sifts to its mathematical setup. Specifically the workflow needs to accommodate the aforementioned suggestions through a joint integrative scheme designed to analyze both data types (GE profiles and network/pathway information). While the full discussion of this process is withheld until Chapter 3, a small preview is presented here. Some of these details include:

- 2D graphs will be used to represent the biological networks and pathways. Furthermore, they also serve as the final data structures corresponding to each individual.

---

[5]Regression based techniques may run the risk of unattainable sample sizes for example. Since the number of features (genes) will most definitely outweigh the number of samples, constructing a saturated model becomes a foregone conclusion. As a result, variations of regression based techniques have often turned to dimension reduction. First a filter is applied. Afterwards, the regression framework will be introduced on the filtered list. This second workflow, while avoiding issues with sample size requirements, is also not free from drawbacks either. Specifically, the filters used to reduce the number of initial features are classic examples of questionable preprocessing techniques required to run the method. While they are sound from a mathematical point of view, they are ultimately just ad-hoc techniques for simplifying a complex biological phenomenon for modeling purposes.

[6]An example approach towards the problem is prediction based on the mean values of input features - i.e. Regression, K-means, etc... While this approach isn't necessarily incorrect, the extreme confidence placed its ability to correctly classify patients is still questionable. After all, what if means aren't the best classifying choice?

- The GE profiles will be infused with the 2D graphs.
- Support vector machine[32] (SVM) will serve as the methodological base.

  Note: While the physical setup of SVM will remain the same, the primary computation involving the kernel function[33] (kernel) will represent innovations presented in this thesis. This unique derivation will consequently permit the specification of additional assumptions and approaches towards the problem all under the presence of non-standard inputs (2D graphs).
- Modeling will be based on both mean- and correlation-based signals. Each approach would be coded by the setup of their 2D graphs and kernel function.

While the ideas presented here are just meant as a prelude to the methodological discussion, it should be evident that the developed response predictor would somehow incorporate SVM, classification based on correlations/means, and 2D graphs all into one logical workflow. As a result the reference used in conjunction to the method would be: '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP).

Since SVM and the idea of positive definite kernels will play a key role in the methodological development of SCRIP, the remainder of this chapter will therefore be reserved for their discussion.

2.3 Support Vector Machine

SVM was developed as a machine learning framework for purposes of supervised binary classification. This popular tool was initially proposed in the 1990s from the statistical learning theory introduced by Vapnik et al.[34] Their close connection with positive definite kernels[35] (kernel functions; kernels) should not come as a surprise as these concepts have become central players in a variety of learning tasks.

Currently applications of SVM can be seen in a variety of fields including multimedia information retrieval[36], pattern recognition[37], and bioinformatics[38]. While this thesis only calls upon its binary classification analogue to model GEP and pathway information, extensions to multiclass classification[39], regression[40], and density estimation[41] also exist. However due to the aforementioned task at hand, the focus here will be to introduce SVM as a linear discriminant function for purposes of binary classification. For a complete presentation of this algorithm, other references[33] are recommended.

2.3.1 Linear Classification

The introduction of SVM starts with the C-classification[42] method reserved for binary prediction. In this classifier, assume that the training instances (two classes in total) are linearly separable. Section 2.3.2 would then relax this requirement.

Formally, given a set of l training objects (i.e. gene expression vectors) $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \in \mathbf{X}$ and their corresponding binary labels (i.e. no-response or response) $\{y_1, \ldots, y_l\} \in \{-1, 1\}$, SVM produces a classifier $f : \mathbf{X} \to \{-1, 1\}$ capable of predicting the class labels of new data instances $\mathbf{x} \in \mathbf{X}$. As mentioned in Section 2.2, since the data inputs of SCRIP will correspond to graphs (representing GE profiles and network/pathway information), the input space $\mathbf{X}$ will therefore reference the graph space $\mathbf{G}$ (the set of all possible labeled graphs) during the application process. For this discussion however, assume that $\mathbf{X}$ is the Eculidian space $\mathbb{R}^p$. Hence the training objects will then reference p-dimensional vectors.

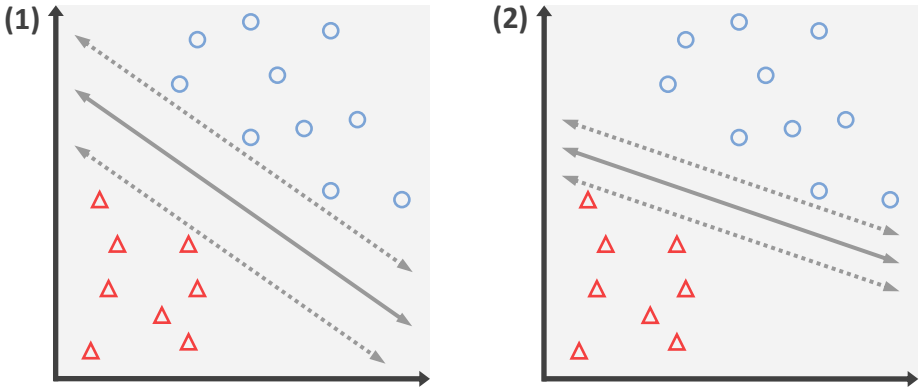


Figure 6: Maximizing Separating Boundary
Assuming that the two classes of points are linearly separable as depicted in (1) and (2), SVM would select the hyperplane that maximizes the distance between the superimposed margins (as denoted by the dotted lines above and below the physical boundary). In the toy example, SVM would therefore default to (1) since it exhibits a wider margin in comparison to (2).

In this vector space, SVM defines a classifier based on the observed interactions between the data points. Consequently this would be formalized as a hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$, where $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Note that since the training vectors are assumed to be linearly separable, there will always exist some $(\mathbf{w}_*, b_*) \in (\mathbb{R}^p \times \mathbb{R})$ (which correspond correspond to a hyperplane) satisfying:

$$y_i \left(\mathbf{w}_*^T \mathbf{x}_i + b_*\right) \geq 0 \; \forall \; i \in \{1, \ldots, l\} \tag{1}$$

As a result a decision function corresponding to a new vector $\mathbf{x}$ can be based on the sign of the corresponding linear function:

$$f(\mathbf{x}) = \text{sign}\left(\mathbf{w}_*^T\mathbf{x} + b_*\right) \tag{2}$$

Geometrically, the chosen hyperplane $\mathbf{w}_*^T\mathbf{x} + b_* = 0$ separates the vector space $\mathbb{R}^p$ into two half halves such that the positive and negative training vectors each lie on distinct sides. However due to the separable assumption, there will exist and infinite number of hyperplanes satisfying the condition (or essentially an infinite number of $(\mathbf{w}_*, b_*)$ pairs that satisfy such condition). As a result SVM will select the hyperplane exhibiting the largest distance between the closest data vector(s) from both classes. This basic rule translates into maximizing the hyperplane margin - a hypothetical distance between two superimposed hyperplanes above and below the original one: $\mathbf{w}_*^T\mathbf{x} + b_* = 1$ and $\mathbf{w}_*^T\mathbf{x} + b_* = -1$ respectively. An illustration of this concept can be seen in Figure 6.

Since the thickness of the margin[7] corresponding to any hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ can be expressed as $2/\left(\sqrt{\mathbf{w}^T\mathbf{w}}\right)$, the optimization process to arrive at the most optimal hyperplane can be expressed in the following minimization (primal form):

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \\ & i = 1, \cdots, l \end{aligned} \tag{3}$$

Note that in this primal form, maximizing the margin is equivalent to minimizing the objective function in Equation 3[8]. The inequality constraints are there to guarantee a hyperplane that correctly classifies all training points. Instances that satisfy the equality however are called the *support vectors* since they are located on one of the two superimposed hyperplanes and metaphorically 'support' the margin (hence the name).

Ultimately, the convex optimization problem in Equation 3 (with a quadratic criterion and linear constraint) usually gets rewritten in a dual formulation since standard quadratic programming techniques can then be applied to the new representation. To therefore arrive at the dual form, Lagrange multipliers[43] are used to re-expressed the problem in an equivalent maximization on $\boldsymbol{\alpha}$ (derivation provided in Section 2.3.4):

---

[7]This margin is in reference to the one generated by superimposing two additional hyperplanes, one above and one below the original one.

[8]Minimize the denominator in order to maximize the entire fraction. Furthermore, since the square root function within the denominator is an increasing function, it can be removed from the optimization procedure.

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & \alpha_i \geq 0 \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ & i = 1, \cdots, l \end{aligned} \tag{4}$$

where $Q_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$; $\mathbf{e} = \{1, \cdots, 1\}^T$.

By solving Equation 4, the solution to the dual problem ($\hat{\boldsymbol{\alpha}} = \{\hat{\alpha}_1, \ldots, \hat{\alpha}_l\}$) will have a unique correspondence with the solution ($\hat{\mathbf{w}}$) of the primal problem (Equation 3). This relationship can be expressed as follows:

$$\hat{\mathbf{w}} = \sum_{i=1}^{l} \hat{\alpha}_i y_i \mathbf{x_i} \tag{5}$$

By plugging the result from Equation 5 into the linear decision function from Equation 2, the resulting decision function ($\hat{f}$) based on the most optimal separating hyperplane ($\hat{\mathbf{w}}^T\mathbf{x} + \hat{b} = 0$) is defined as follows:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \text{sign}\left(\hat{\mathbf{w}}^T\mathbf{x} + \hat{b}\right) \\ &= \text{sign}\left(\left[\sum_{i=1}^{l} \hat{\alpha}_i y_i \mathbf{x_i}\right]^T \mathbf{x} + \hat{b}\right) \\ &= \text{sign}\left(\sum_{i=1}^{l} \hat{\alpha}_i y_i \mathbf{x_i}^T \mathbf{x} + \hat{b}\right) \end{aligned} \tag{6}$$

2.3.2 Linear Classification With Soft Margin

Since the basic C-classification method is only valid for linearly separable data, it becomes too restrictive in many real life applications. Therefore to relax the later part of this assumption (to account for linearly inseparable), a slight modification to the optimization process (primal and dual forms) is required. This is achieved by introducing the concept of a 'soft margin[44]'.

The idea of a soft margin allows mislabeled vectors to still be classified. This extension is necessary since a hyperplane capable of splitting all training examples will no longer exist under this scenario. As a result the soft margin method will resort to the hyperplane that conducts a split as cleanly as possible - while still maximizing the distance to the nearest cleanly split examples. In order to carry out this optimization, the primal and dual forms are relaxed using slack variables ξ_i, which measure the degree of misclassification of each training vector $\mathbf{x}_i$. In particular, the primal problem introduced in Equation 3 will be modified as follows (modifications are highlighted in red):

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i \\ & 0 \leq \xi_i \\ & i = 1, \cdots, l \end{aligned} \tag{7}$$

, where the introduction of C is used to control the trade off between training errors and rigid margins. In other words, increasing the value of C will force a more accurate model (better training error) that may unfortunately not generalize well. The corresponding modification to the dual problem is as follows (modifications are highlighted in red):

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ & i = 1, \cdots, l \end{aligned} \tag{8}$$

where $Q_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$; $\mathbf{e} = \{1, \cdots, 1\}^T$.

Note that even though a modified version of the dual problem (Equation 8) is solved, the relationship specified in Equation 5 will still hold. Furthermore since the linear decision function remains the same, Equation 6 will also hold.

2.3.3 Kernel Classification With Soft Margin

Even though the previous formulation of SVM allows for misclassification, it is still too restrictive since linear classifiers (optimal hyperplane in the original feature space) resemble the only options available to train the data. To therefore relax this final assumption, SVM uses a 'kernel function' so that generalized decision surfaces can be obtained.

A kernel function effectively enables training vectors to be mapped into higher dimensions where hyperplanes can again facilitate as the separator. While this concept would seem similar to the previous formalization, the distinction lies in the mapping and its ability to re-represent a linear boundary in the altered space as a nonlinear one in the original. Due to this concept, it omits the need to fit nonlinear curves to the training data which will consequently require drastic changes to the existing optimization theory. Instead only slight modifications to the training data will be required. Specifically the requirements are the map Φ and its corresponding mapped space $\mathscr{H}$:

$$\begin{aligned} \Phi : \mathbb{R}^p &\rightarrow \mathscr{H} \\ \mathbf{x} &\rightarrow \Phi(\mathbf{x}) \end{aligned} \tag{9}$$

In reference to both of these components, $\mathscr{H}$ usually carries additional significance since it defines the mapping function used in conjunction to the alternative space. Therefore in an application framework it represents the only piece that requires specification. And in particular since few restrictions are placed on its form, it is logically easy to understand despite the complications associated with its actual implementation.

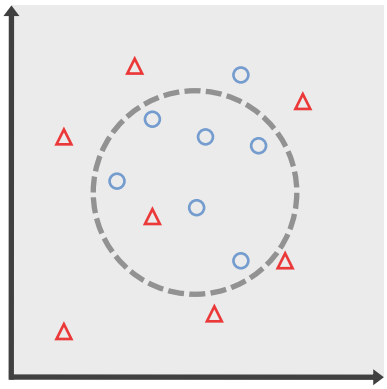


Figure 7: Nonlinear Boundaries

In certain cases, a linear boundary would no longer conduct the most optimal classification. Therefore the flexibility to impose a non-linear boundary becomes a necessary option in order to guarantee the efficacy associated with the classifier.

Nevertheless the only requirement concerning $\mathscr{H}$ is the presence of a legitimate inner product. In other words $\mathscr{H}$ needs to be a Hilbert space[45] so that $\Phi(\mathbf{x})^T\Phi(\mathbf{x}')$ exists for any two vectors $\mathbf{x}, \mathbf{x}'$ defined in $\mathbb{R}^p$. And in practice since the data will only be used in reference to this formulation, which coincidentally also defines $\mathscr{H}$, it therefore becomes sufficient to construct the inner product as a means to satisfy all conditions. This sufficiency can be summarized as the kernel trick[46]:

$$k(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T\Phi(\mathbf{x}') \;\forall\; \mathbf{x},\mathbf{x}' \in \mathbb{R}^p \tag{10}$$

where $k(\mathbf{x},\mathbf{x}')$ references the kernel function.

Since a classical result states that any function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ can be plugged into SVM given its symmetry and positive definiteness, the corresponding changes involving soft margins (Equation 7) will be as follows (modifications are highlighted in red):

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i\left(\mathbf{w}^T\Phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i \\ & 0 \leq \xi_i \\ & i = 1,\cdots,l \end{aligned} \tag{11}$$

And the changes to the corresponding dual problem will be as follows (modifications are highlighted in red):

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha} \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \\ & \mathbf{y}^T\boldsymbol{\alpha} = 0 \\ & i = 1, \cdots, l \end{aligned} \tag{12}$$

where $Q_{ij} = y_i y_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$; $\mathbf{e} = \{1, \cdots, 1\}^T$.

By solving Equation 12, the optimal solution ($\hat{\boldsymbol{\alpha}}$) will have the following relationship with the optimal solution ($\hat{\mathbf{w}}$) from Equation 11:

$$\hat{\mathbf{w}} = \sum_{i=1}^{l} \hat{\alpha}_i y_i \Phi(\mathbf{x_i}) \tag{13}$$

The resulting non-linear decision function will then become:

$$\hat{f}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{l} \hat{\alpha}_i y_i k(\mathbf{x_i}, \mathbf{x}) + \hat{b}\right) \tag{14}$$

2.3.4 Linear Classification Primal To Dual

The dual representations of SVM, in reference to the primal and dual problems from Sections 2.3.1 to 2.3.3, were introduced for the sole purpose of computation. Since the optimization procedure described in these problems can potentially involve high dimensions, SVM takes advantage of the kernel trick and its dual formulation to avoid these issues.

In the linear classification case for example, the constrained quadratic problem (Equation 3 of Section 2.3.1) has a dimension size of $p+1$ ($\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$). In this primal setup optimization is therefore only guaranteed under a small p, i.e. $\leq 10^3$. For larger values of p the curse of dimensionality takes over making outright optimization an impractical option. Fortunately due to the Kuhn-Tucker theorem[47] and the corresponding convexity associated with Equation 3, Lagrange multipliers can be used to transform the above problem into its equivalent dual form. Under this new formulation the variables are subjected to looser constraints and thus optimization, regardless of the dimensionality of the primal setup, becomes feasible.

In the remainder of this discussion, the derivation for the dual problem is provided. Although this particular derivation is restricted to a linear classifier, the same logic can be extended to account for the soft margin and/or kernel

function. However due to the lengthy arguments required, only the first example is shown. In particular, to transform Equation 3:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \\ & i = 1, \cdots, l \end{aligned}$$

, from Section 2.3.1 into its corresponding dual form, the first step is to compute the associated Lagrange function ($\boldsymbol{\alpha}$ corresponds to the Lagrange multipliers):

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{l} \alpha_i \left(y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right) \\ &= \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{l} \alpha_i \left(y_i\mathbf{w}^T\mathbf{x}_i - 1\right) - b\sum_{i=1}^{l} \alpha_i y_i \end{aligned} \tag{15}$$

In the framework of this particular optimization problem, strong duality[48] holds. Thus the dual form is obtained by:

$$\text{Dual Form} = \max_{\boldsymbol{\alpha} \geq 0} \left(\min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \right) \tag{16}$$

To solve Equation 16, the first step will be to simplify the entire expression by consider the inner minimization by itself. Thus for some fixed $\boldsymbol{\alpha}$ the minimization becomes:

$$\min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = \begin{cases} -\infty & \text{if } \sum_{i=1}^{l} \alpha_i y_i \neq 0 \\ \min_{\mathbf{w}} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{l} \alpha_i \left(y_i\mathbf{w}^T\mathbf{x}_i - 1\right) \right\} & \text{if } \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \tag{17}$$

Since b is unrestricted the first condition follows by setting $b = \pm\infty$ when $\sum \alpha_i y_i \neq 0$. Thus depending on the sign of the summation, the minimization is automatically achieved at $-\infty$ regardless of $\mathbf{w}$. When $\sum \alpha_i y_i = 0$, the minimization occurs when:

$$\begin{aligned} & \frac{d}{d\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{l} \alpha_i \left(y_i\mathbf{w}^T\mathbf{x}_i - 1\right) \right) = 0 \\ \Rightarrow \hat{\mathbf{w}} &= \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \end{aligned} \tag{18}$$

Thus the second condition in Equation 17 becomes:

$$\begin{aligned}
\min_{\mathbf{w}} \left\{ \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{l} \alpha_i \left(y_i \mathbf{w}^T \mathbf{x}_i - 1 \right) \right\} &= \frac{1}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} - \sum_{i=1}^{l} \alpha_i \left(y_i \hat{\mathbf{w}}^T \mathbf{x}_i - 1 \right) \\
&= \frac{1}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} - \hat{\mathbf{w}}^T \left[\sum_{i=1}^{l} \alpha_i \left(y_i \mathbf{x}_i \right) \right] + \sum_{i=1}^{l} \alpha_i \\
&= \frac{1}{2}\hat{\mathbf{w}}^T\hat{\mathbf{w}} - \hat{\mathbf{w}}^T\hat{\mathbf{w}} + \sum_{i=1}^{l} \alpha_i \\
&= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \left[\sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i \right]^T \left[\sum_{j=1}^{l} \alpha_j y_j \mathbf{x}_j \right] \\
&= \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \qquad (19)
\end{aligned}$$

By plugging the result from Equation 19 into Equation 17, the original maximization in Equation 16 becomes:

$$\max_{\boldsymbol{\alpha} \geq 0} \left(\min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \right) = \begin{cases} \max_{\boldsymbol{\alpha} \geq 0}(-\infty) & \text{if } \sum_{i=1}^{l} \alpha_i y_i \neq 0 \\ \max_{\boldsymbol{\alpha} \geq 0} \left(\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) & \text{if } \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \qquad (20)$$

Since the final goal is to perform maximization over $\boldsymbol{\alpha}$, $-\infty$ is definitely not the answer. As a result the dual form defaults to the second condition from Equation 20. In this formulation, $\sum \alpha_i y_i = 0$ subsequently becomes the constraint.

References

[1] F. Bertucci, *et al.*, *Oncogene* **23**, 1377 (2004).

[2] Y. Wang, *et al.*, *Journal of Clinical Oncology* **22**, 1564 (2004).

[3] K. Iwao-Koizumi, *et al.*, *Journal of clinical oncology* **23**, 422 (2005).

[4] M. Jansen, *et al.*, *Journal of Clinical Oncology* **23**, 732 (2005).

[5] L. van't Veer, *et al.*, *nature* **415**, 530 (2002).

[6] K. Iwao, *et al.*, *Human molecular genetics* **11**, 199 (2002).

[7] A. Anguiano, *et al.*, *Journal of Clinical Oncology* **27**, 4197 (2009).

[8] O. Decaux, *et al.*, *Journal of Clinical Oncology* **26**, 4798 (2008).

[9] J. Moreaux, *et al.*, *Haematologica* **96**, 574 (2011).

[10] F. Zhan, *et al.*, *Blood* **108**, 2020 (2006).

[11] B. Ghadimi, *et al.*, *Journal of Clinical Oncology* **23**, 1826 (2005).

[12] E. Jensen, J. McLoughlin, T. Yeatman, *Current opinion in oncology* **18**, 374 (2006).

[13] J. Mariadason, *et al.*, *Cancer research* **63**, 8791 (2003).

[14] K. Nagasaki, Y. Miki, *Breast Cancer* **15**, 117 (2008).

[15] M. Schauer, *et al.*, *Clinical Cancer Research* **16**, 330 (2010).

[16] D. Ross, *et al.*, *Nature genetics* **24**, 227 (2000).

[17] U. Scherf, *et al.*, *Nature genetics* **24**, 236 (2000).

[18] J. Torres-Roca, *et al.*, *Cancer research* **65**, 7169 (2005).

[19] A. Potti, *et al.*, *Nature medicine* **12**, 1294 (2006).

[20] P. Lønning, S. Knappskog, V. Staalesen, R. Chrisanthar, J. Lillehaug, *Annals of oncology* **18**, 1293 (2007).

[21] R. Luthra, *et al.*, *Journal of clinical oncology* **24**, 259 (2006).

[22] M. Del Rio, *et al.*, *Journal of clinical oncology* **25**, 773 (2007).

[23] L. Van't Veer, R. Bernards, *Nature* **452**, 564 (2008).

[24] T. Sørlie, *et al.*, *Molecular cancer therapeutics* **5**, 2914 (2006).

[25] J. Mariadason, D. Arango, L. Augenlicht, *Drug resistance updates* **7**, 209 (2004).

[26] R. Bast Jr, *et al.*, *Journal of Clinical Oncology* **19**, 1865 (2001).

[27] D. Albertson, C. Collins, F. McCormick, J. Gray, *et al.*, *Nature genetics* **34**, 369 (2003).

[28] M. Kanehisa, *et al.*, *Novartis Foundation symposium* (Chichester; New York; John Wiley; 1999, 2002), pp. 91–100.

[29] D. Nishimura, *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117 (2001).

[30] G. Joshi-Tope, *et al.*, *Nucleic acids research* **33**, D428 (2005).

[31] G. Milne, M. Nicklaus, J. Driscoll, S. Wang, D. Zaharevitz, *Journal of chemical information and computer sciences* **34**, 1219 (1994).

[32] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[33] A. Smola, B. Schölkopf, *Learning with kernels* (Citeseer, 1998).

[34] V. Vapnik, *The nature of statistical learning theory* (springer, 1999).

[35] J. Shawe-Taylor, N. Cristianini, *Kernel methods for pattern analysis* (Cambridge university press, 2004).

[36] Y. Cao, *et al.*, *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (ACM, 2006), pp. 186–193.

[37] C. Burges, *Data mining and knowledge discovery* **2**, 121 (1998).

[38] E. Byvatov, G. Schneider, *et al.*, *Applied bioinformatics* **2**, 67 (2003).

[39] C. Hsu, C. Lin, *Neural Networks, IEEE Transactions on* **13**, 415 (2002).

[40] A. Smola, B. Schölkopf, *Statistics and computing* **14**, 199 (2004).

[41] S. Mukherjee, V. Vapnik, *Center for Biological and Computational Learning. Department of Brain and Cognitive Sciences, MIT. CBCL* (1999).

[42] C. Hsu, C. Chang, C. Lin, *et al.*, A practical guide to support vector classification (2003).

[43] R. Bellman, *Proceedings of the National Academy of Sciences of the United States of America* **42**, 767 (1956).

[44] K. Veropoulos, C. Campbell, N. Cristianini, *et al.*, *Proceedings of the International Joint Conference on Artificial Intelligence* (Citeseer, 1999), vol. 1999, pp. 55–60.

[45] A. Gleason, *The Logico-algebraic Approach to Quantum Mechanics: Historical evolution* **5**, 123 (1975).

[46] B. Scholkopf, *et al.*, *Advances in neural information processing systems* pp. 301–307 (2001).

[47] M. Hanson, *Journal of Mathematical Analysis and Applications* **80**, 545 (1981).

[48] M. Ramana, L. Tunçel, H. Wolkowicz, *SIAM Journal on Optimization* **7**, 641 (1997).

Methodological Development Of SCRIP

Chapter 3

This chapter details the methodological development of '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP). The discussion starts with its motivation and reference to SVM. Afterwards the workflow and mathematical details will be presented. The later part of this discussion will include a thorough interpretation of the graph structures involved, kernel computation, preprocessing techniques of 'overall graphs', and merge process between GE/pathway information. Again, as constructed, SCRIP will only utilize the GE data. Chapter 6 deals with its expansion to CN data.

3.1 Introduction

'**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) is developed as an integrative tool for response prediction. Its design is predicated on the shortcomings of current methods many of which are outlined in Section 2.2. By resolving some of these issues, improvements in predictive reliability and accuracy can hopefully be achieved.

From an integrative point of view, SCRIP merges pathway information with the standard GE data. The selection of such data type (pathway information), as opposed to any other source of genomic information, is a revelation coming from two different angles. First, since response is dictated by a variety of mechanisms[1], adding in an additional layer of covariates only helps to solidify the predictive platform. In comparison to other sources of genomic information, pathways are the ideal since they allowed gene relationships to be interrogated at a group level[2]. Hence a more comprehensive gene-based treatment (that still improves the predictive platform) will be automatically adopted upon their inclusion. Second, since response based studies are often restricted to GE profiles, the integrated data type must therefore be available without additional experimental work or else the practicality and relevance of the developed method will be compromised. Thus while other genomic sources better suited for response prediction exist, the infrequency at which they appear makes their inclusion an implausible proposition. On the other hand since pathways can be easily accessed through online databases, designating them as the integrated target becomes an obvious choice.

With the inclusion of pathway information, SCRIP then looked to adopt a flexibility methodological base under which: (1) Data integration is feasible; and (2) Multiple assumptions and approaches towards classification are allowed. In no particular order of satisfying these requirements, the theory of positive definite kernels[3] and support vector machines[4] (SVM) was called upon to alleviate these issues.

In particular SVM serves as an ideal methodological base since it can easily generalize to nonvectorial spaces[9]. With respect to the aforementioned requirements, this property has resounding consequences. First, it allows straightforward extension of linear SVM (Section 2.3.1) to a nonlinear setting without loosing the advantages inherent to the original formulation (unicity of the solution, robustness to overfitting, etc...). And second, given that the corresponding kernel can be defined, it enables application to nonvectorial data without being restricted by a predetermined approach.

---

[9]The nonvectorial space, say the set of all 2D graph structures, will then need to be embedded into a vector space through some mapping function. Afterwards, linear SVM can be applied to the training points since a kernel function is defined. These concepts refer to the 'Kernel Classification With Soft Margin' ideas introduced in Section 2.3.3.

These properties consequently infuse SVM with the required flexibility[10] without loss of robustness or practicality.

To demonstrate the proposed flexibility, consider response signatures predicated on the means of expression values. In this standard approach each response group is assumed to follow a unique expression pattern amongst a set of genes. Under the validity of this biological hypothesis, classification implies using expression vectors as the data representations and subsequently an inner product defined between these vectors as the kernel function (regular dot product, polynomial ,or radial-basis kernel). As a result SVM metaphorically compares the expression values through the designated kernel to thereby generate a prediction model. Upon completion the designated approach is automatically carried out through the specification of these two parameters.

Not surprisingly SCRIP will also take advantage of this aforementioned flexibility to implement its unique approach. Starting with the biological hypothesis, it assumes that each response group is dictated by a unique 'correlation and mean' based signature. Consequently the combined pairwise correlation patterns between genes along with their physical expression values will be used for modeling; An innovation that while foreign to response prediction, has witnessed success in other statistical frameworks thus suggesting merit for their combination[5]. And finally on top of this, pathways will be integrated into the workflow to accompany the gene expression data. The details on the approach and integration step will be discussed in Section 3.2.2.

In reference to the corresponding methodological setup, the implementation of the approach boils down to the kernel function. While this discussion is postponed until Section 3.6, SCRIP will ultimately employ labeled graphs as the data representations so that the ideas of correlations, means, and pathways can all be seamlessly merged together. Here these graphs will receive the following notation: Vertices represent genes; Edges represent correlation-based relationships derived from the physical gene expression values. As a result of this representation, an inner product defined between labeled graphs would be required for the analysis, comparison, and classification of such data structures. This naturally brings up the concept of 'graph kernels[6]'.

In particular, a variety of graph kernels have been proposed during this past decade. Amongst them the approach involving vectoral mappings of the input graphs has become the mainstream formulation[7]. Graph similarity as attested in these methods is simply defined as the dot product between the corresponding vectors. Consequently the construction of these kernels will solely rest on the mapping used. With SCRIP, the implemented kernel function will

[10] Allows analysis according to multiple assumptions and non-standard data input.

feature a map using 'walks' as the comparison base[6]. In reference to this mapping, the corresponding elements of the vector will then designate the number of times a predetermined walk can be performed. Thus the resulting dot product between such vectors will evaluate graph similarity based on these shared pairwise features. Additional details on these concepts will be presented later in Sections 3.6.1 and 3.6.3.

In conclusion, SCRIP presents an alternative approach for response prediction that attempts to alleviate some of the current deficiencies. First, instead of mimicking traditional classification techniques predicated on means, SCRIP also emphasized pairwise correlation patterns. And second, by integrating pathway information into the computational workflow, the inability of GE to completely predict response can be partially accounted for. By implementing the expanded approach, the hope is that a desired signature can finally be obtained.

In the remainder of this chapter, a complete discussion of the methodological development behind SCRIP will be provided. Here the novelties, workflow, graph details, kernel construction, and merge process inherent to SCRIP will be presented. Subsequent chapters will then deal with the simulation and application process.

3.2 Workflow Of SCRIP

SCRIP adopts a biological hypothesis that associates a unique correlation and mean based signature with a response group. Here the 'mean' part of the signature refers to the classical approach of classification based on the physical expression values[8]. Since this idea has been thoroughly explored, the following workflow along with the methodological discussion (Sections 3.2 to 3.6) will be heavily tailored to the correlation part of the design. The subsequent discussion of the overall graph, individual graph, merge method, and kernel will also represent novelties introduced on its behalf.

3.2.1 Workflow For Mean

The mean part of SCRIP was adopted from the classical approach of model fitting. First, the initial list of input features (genes) will be filtered[9] based on a predefined criterion[11]. Afterwards this filtered list will then be used

[11]The filtering process can be defined in a variety of ways depending on the preference of the user. Variance and mean filters are commonly used. In this thesis, filters based on the overall graph are also considered. Here the overall

for training purposes. Since the training conducted here is predicated on means, the setup within SVM will involve vectors as data representations and a corresponding kernel to compute their inner product[10] (i.e. Linear, polynomial, or radial basis kernel). Once this entire procedure is carried out, predictions were subsequently obtained. The combination of these results with the correlation output will then form the final predictions of SCRIP.

3.2.2 Workflow For Correlation

In reference to the correlation part of the signature, SCRIP will use 'correlation' and 'co-expression' (between a gene pair) interchangeably. The differentiation is only made for purposes of explanation as correlations are better suited with the computation while co-expressions have greater appeal in the context of biology. At the end of the day however, by defining co-expression as two genes exhibiting similar expression values, they referred to the same concepts. Thus a high (positive) correlation $(\rho \geq 0.8)$ between two genes will indicate the presence of the co-expression relationship and a low correlation $(|\rho| \leq 0.05)$ will indicate otherwise.

With the proposed setup, a toy example of these co-expression signatures can be seen in (1) of Figure 8. Here the response specific co-expression signatures are portrayed as graphs where the nodes represent the genes and the (present/absent) edges represent (existence/absence of) co-expression relationship. For example the response group co-expression signature is indicated by the co-expression (or high correlation) between genes A-B and B-C. Similarly the non-response group co-expression signature is indicated by genes A-B and B-D. These unique co-expression signatures will be referred to as the 'response specific graphs' (RSG).

While the idea of RSGs effectively summarized the co-expression relationships unique to each group, they are unfortunately just abstractions of the underlying biology and therefore hidden. Nevertheless while access to each RSG is off limits, SCRIP assumes that their combined information is embedded within the concept of an 'overall graph' (OG). Here OGs are defined similar to a biological prior that pinpoints a set of gene pairs whose co-expression patterns can directly impact response. In other words the OG reveals all 'important' gene pairs without identifying their response specific configurations. Thus in this context an 'important' gene pair indicates the possibility that the co-expression status (co-expressed vs. not co-expressed) can be different between responders and non-responders. A complete discussion of the OG will be presented in Section 3.3 and the toy example corresponding to the previously mentioned RSGs is shown in (2) of Figure 8.

---

graph acts like a biological prior that designates a group of interesting genes through the nodes.

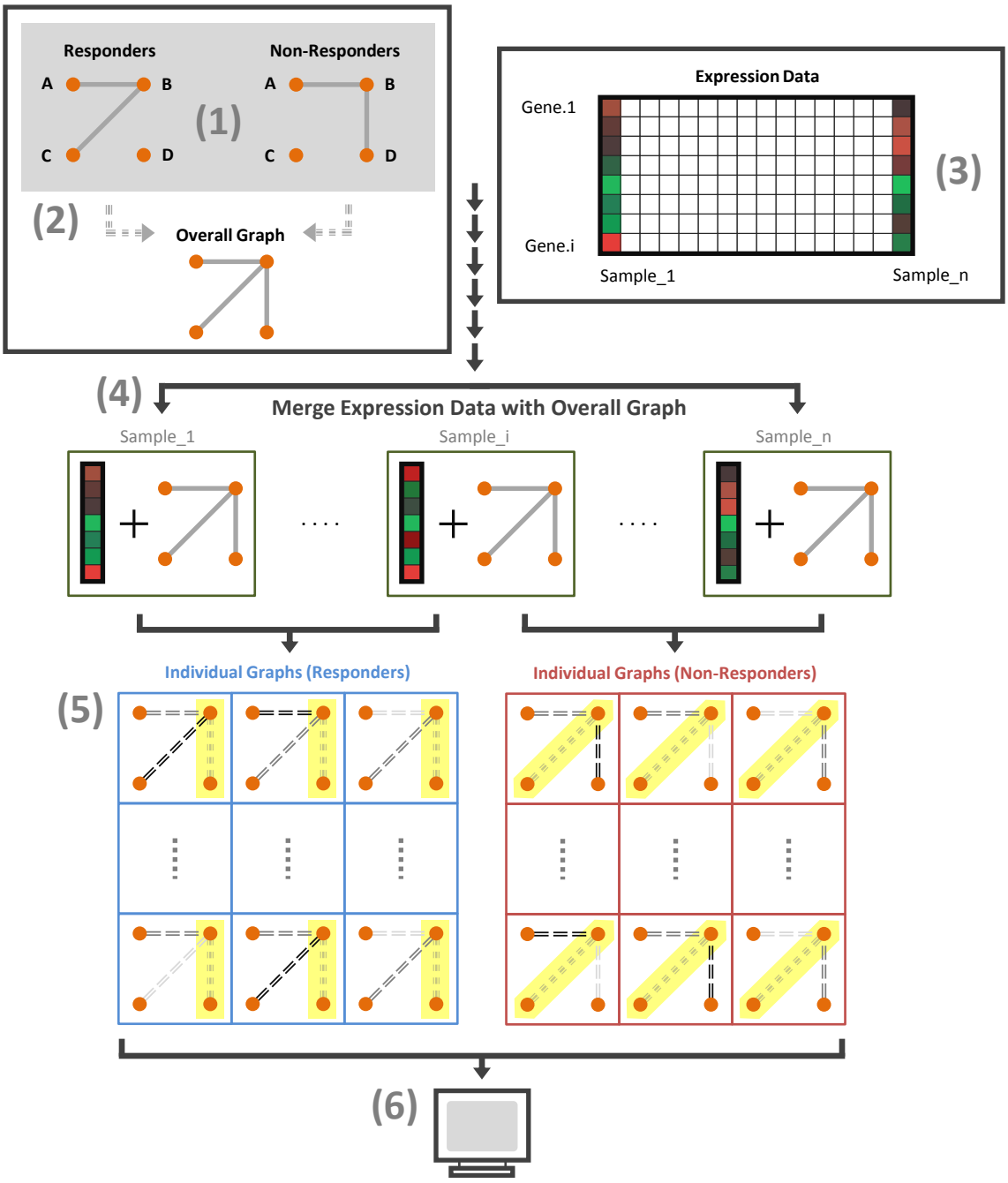


Figure 8: Workflow Of Correlation Signature

The OG pinpoints all informative co-expression relationships but doesn't reveal their unique configuration within each response group. Thus for any linked gene pair, there exists the possibility that the status (co-expression exists vs. absent) is different between responders and non-responders. The merge method therefore examines all linked gene pairs within an individual and for each given instance, assigns probabilistic membership information. The membership information will then be processed by SVM to complete the algorithm.

Since the make-up of an OG closely resembles a treasure map that locates all potentially important gene pairs, deciding on what constitutes as a candidate OG(s) therefore becomes a free choice. While this privilege increases the flexibility associated with SCRIP, it however comes at a cost. On one hand, while any type of gene-map involving edges and nodes can qualify, the questionable effectiveness of these seemingly arbitrary designations becomes an obstacle for the analysis[12]. Thus while it is easy to designate OGs, the true difficulty becomes verifying its ability to capture response specific co-expression patterns.

Nevertheless assuming that this previous point can be resolved and that a candidate OG can be selected, SCRIP will then 'merge ((4) of Figure 8)' the expression data ((3) of Figure 8) with this prior to arrive at the individual graphs (IG) ((5) of Figure 8). Briefly IGs are meant as a reflection of an individual's 'membership status' with respect to the important gene pairs. Since a gene pair is assumed to follow a mixture of two bivariate distributions (**H** - high correlation; **L** - low correlation), membership status therefore corresponds to a probabilistic label ($P(\text{from } \mathbf{H})$ or $P(\text{from } \mathbf{L})$) indicating the likelihood of **H** as the generator (to the given observation pair). If the truth is **H**, as attested by a link in the RSG, then the collective membership status amongst these individual should be higher in comparison to the group where the truth is **L**. It is exactly due to this reasoning that an effective OG (one which contains the differential information as depicted in Figure 8) can produce membership statuses that are consistent within each response group yet different across. Consequently such mixture of consistency and conflict will then enable classification predicated on these signals. Additional details on the IGs and merge will be presented later in Sections 3.4 and 3.5.

Returning back to the workflow, once the IGs are obtained, SCRIP will then define a (graph) kernel so that the analysis of such representations becomes possible. Due to the complimentary nature of kernels and support vector machines (SVM), SCRIP will ultimately generate the correlation part of the prediction model by ushering into the avenue opened by SVM ((6) in Figure 8). Additional details on the kernel construction will be presented later in Section 3.6.

SCRIP differs from classical approaches since response prediction is predicated on co-expression- and mean-based

---

[12]From a methodological standpoint, SCRIP will require at least one effective OG so that classification based on co-expression is feasible. While such requirement may place an extreme amount of pressure behind the formulation of these priors, it is offset by no restriction that caps their starting number. In other words, as long as the starting batch of OGs contains an effective one, SCRIP should theoretically produce a desirable classifier. In regards to the actual application, an informed shotgun approach is used to generate this initial batch. Specifically, pathways/networks from online databases (KEGG, BioCarta, Reactome, and NCI) are assembled and then screened by iteratively applying SCRIP to each input. Afterwards, the selected subset will be used to construct the final model. An in dept discussion regarding this process will be presented later in Chapter 5.

patterns. By introducing networks and pathways into the computation, it also takes advantage of a biologically sound weighing scheme that naturally integrated into the SVM algorithm.

3.3 Overall Graphs

The overall graph (OG) resembles a biological prior that pinpoints a set of important gene pairs whose co-expression relationships play an important role in determining response. Here an important gene pair indicates the possibility that the co-expression status (present or absent) can be different amongst responders and non-responders. Following the presentation in this current chapter, this particular set of gene pairs can also be called the 'designated gene pairs' as indicated by edges in the OG.

A distinctive property of the OGs is their 'inclusive' yet 'non-revealing' nature. In other words they are inclusive since all important co-expression relationships are noted and non-revealing since the exact configuration (co-expressed or not co-expressed) within each response group is hidden. As a result of such relationship, there exists the possibility that a designated gene pair will exhibit the same co-expression status amongst responders and non-responders alike. In this case the designated gene pair is 'non-informative[13]'. With such setup, designation therefore only confirms the possibility of a differential co-expression status. By all means it isn't an indicative statement that validates the difference. However in the event that they are (co-expressed in one group and absent in the other), then the designated gene pair becomes 'informative'. Thus all informative gene pairs will have opposite co-expression statuses and can therefore be used to differentiate between the response groups.

SCRIP will assume that an 'effective' OG pinpoints AT LEAST ONE informative gene pair amongst the set of designated ones. An effective OG therefore enables classification based on its signals.

[13] A missing edge between two genes ALWAYS implies that the co-expression status for that particular gene pair is the same regardless of response group. In those cases the gene pair is also considered non-informative. While non-informative gene pairs may still be biologically relevant, their impact on response is assumed to be minimal under the working hypothesis and hence are omitted from consideration.

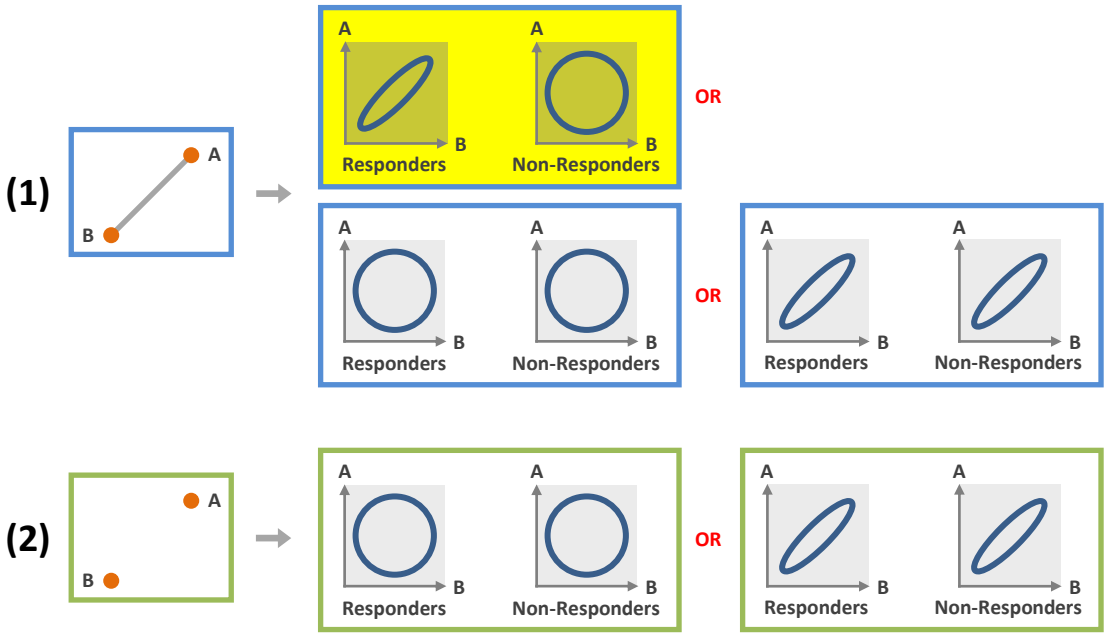


Figure 9: Edges Of The Overall Graph

A designated gene pair (1) within the OG implies the possibility that the correlation status amongst the responders and non-responders would be different. Thus a scenario similar to the blue box highlighted in yellow could happen. However in the other case where the gene pair isn't designated (2), then the correlation status amongst responders and non-responders would be consistent.

3.4 Individual Graphs

While the overall graph (OG) pinpoints a set of important gene pairs, each individual graph (IG) reflects the 'membership status' of a given individual with respect to this information. Here the membership status is constructed as a direct consequence of the assumptions placed behind these features. Specifically for a given gene pair (X,Y), whose expression levels are assumed to follow some distribution, the following properties were adopted:

- In the population where all individuals come from, (X,Y) follows a mixture of two bivariate distributions:

  H: High correlated bivariate distribution;

  L: Low correlated bivariate distribution;

  The mixing proportion of **H** and **L** are $p_1 : p_2$ respectively. They can be 0:1 or 1:0.

- An instance from the high correlated distribution implies $(X,Y) \sim \mathbf{H}$.

  Thus (X,Y) will have a co-expression relationship present in a subset of individuals (proportion of p_1 of all individuals).

- An instance from the low correlated distribution implies $(X,Y) \sim \mathbf{L}$.

Thus (X,Y) will have no co-expression relationship present in a subset of individuals (proportion of p_2 of all individuals).

- Given an observation from any gene pair, only ONE of **H** or **L** can generated it.

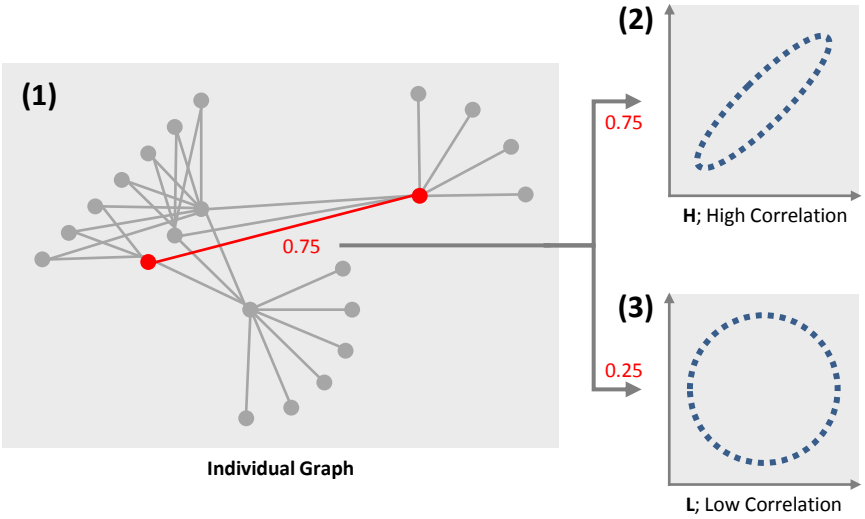


Figure 10: Visualization Of Individual Graphs
All links in the IG (1) are associated with probabilities representing the likelihood that H, high correlated distribution (2), generated the corresponding expression pair for that individual. This probability is a conjugate to the likelihood that L, low correlated distribution (3), generated it. Highlighted link: the individual's observed expression for that highlighted gene pair is more likely from H (75% confidence) than from L (25% confidence).

From these assumptions, the membership status corresponding to each gene pair observation (x,y) is defined as a probabilistic label indicating the likelihood of **H** as the generator. In other words it represents an assignment[14] of the individual's underlying distribution with respect to that gene pair. Thus edges associated with higher probabilities (higher membership status) will indicate greater confidence in **H** as the underlying distribution for the expression instance; while lower probabilities will indicate **L**. Note that the calculated membership status is a conjugate to **L** since only one of **H** or **L** could have generated (x,y).

Collectively an individual's membership status corresponding to all of the designated gene pairs will form their IG. SCRIP hypothesizes that individuals within the same response group will exhibit similar IGs and hence a particular

[14]These assignments are based on probabilities.

configuration of membership statuses[15]. By taking advantage of these similarities, SCRIP will then be able to assign the most likely response label corresponding to a new individual through a comparison of these features.

3.5 Merge Method

With the full discussion of the overall graph (OG) and individual graph (IG) in place, the merge method then bridges these concepts with the GE data. In particular it will use the OG and corresponding vector of expression values to guide the computation of each IG.

As a reminder, links in the IG correspond to probabilistic membership labels for that particular gene pair. Since these pairwise observations can only be an instance from a high (**H**) or low (**L**) correlated distribution, membership status therefore defines the likelihood of the underlying generator amongst these two potential targets. As the computation process behind this decision, the merge method will initially calculate the probability associated with either choice: $p((x,y) \text{ from } \mathbf{H})$ and $p((x,y) \text{ from } \mathbf{L})$ where (x,y) represents the observation. Afterwards $p((x,y) \text{ from } \mathbf{H})$ will be coupled with the edge. To help facilitate this computation, additional assumptions were adopted:

- In the population where all individuals come from, the expression values for each gene follows a normal distribution;
- In the population where all individuals come from, the expression values for each gene pair follows a mixture of two bivariate normal distributions:
  H: High correlated standardized[16] bivariate normal with correlation coefficient $\rho = 0.85$;
  L: Low correlated standardized bivariate normal with correlation coefficient $\rho = 0.05$;
  The mixing proportion of **H** and **L** are $p_1 : p_2$ respectively. They can be 0:1 or 1:0.

To obtain the complete membership information (or the entire IG), first consider the case when inference is restricted to only one gene pair from a particular individual. In this example, the merge method will initially calculate the probability that this observation (x,y) comes from **H** and **L** respectively. Since these probabilities reflect the likeliness of either distribution as the generator, they can be used to construct the membership status assigned to each link.

---

[15]This follows from the differential RSGs.

[16]The standardized bivariate normal takes $\sigma_1 = \sigma_2 = 1$ and $\mu_1 = \mu_2 = 0$

Specifically, SCRIP will create these labels through a comparison of the likelihoods previously computed. The idea here is simple; whichever likelihood is larger should ideally highlight the generator. Thus:

- $p\left((x,y) \text{ from } \mathbf{H}\right) = \frac{p((x,y)|\mathbf{H})}{p((x,y)|\mathbf{H})+p((x,y)|\mathbf{L})}$
- $p\left((x,y) \text{ from } \mathbf{L}\right) = \frac{p((x,y)|\mathbf{L})}{p((x,y)|\mathbf{H})+p((x,y)|\mathbf{L})}$

Fortunately this formulation of the membership matches the aforementioned objective. If (x,y) is indeed an instance of $\mathbf{H}$, then $p\left((x,y) \text{ from } \mathbf{H}\right)$ will be larger than $p\left((x,y) \text{ from } \mathbf{L}\right)$ in a majority of such observations. Likewise if the expression pair is an instance of $\mathbf{L}$, then $p\left((x,y) \text{ from } \mathbf{L}\right)$ will be larger.

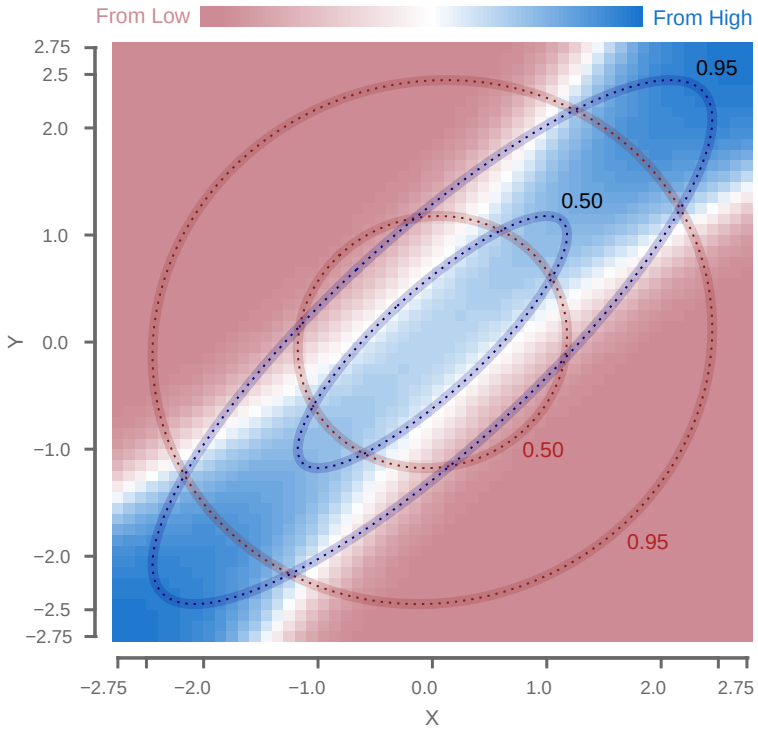


Figure 11: Intuition Behind Merge Process

Heatmap represents the likelihood ratio between a high ($\rho = 0.85$; $\mathbf{H}$) and low ($\rho = 0.05$; $\mathbf{L}$) correlated bivariate normal. The Cartesian plane depicts combinations of expression values and their corresponding probabilistic label (membership information) calculated from the merge. Red: point is more likely from $\mathbf{L}$. Blue: point is more likely from $\mathbf{H}$. Dotted ovals: CDF contour lines for $\mathbf{L}$ (red ovals) and $\mathbf{H}$ (blue ovals) (i.e. 50% of all points from $\mathbf{H}$ lie within the inner blue oval).

To finally obtain the entire IG[17] corresponding the individual's expression vector, the described merge method will

[17] Also the complete membership information.

be repeatedly applied to each designated gene pair in the OG. Therefore at the end of this process, the IG will support the same configuration of edges and nodes as the OG. The only difference however will be the added membership statuses corresponding to each edge.

3.6 Kernel Construction

Before introducing the kernel used in reference to the individual graphs (IG), a secondary interpretation of the edges and associated probabilities will be necessary so that entire design can account for their presence. While previously defined in terms of membership status [18] for a particular gene pair, they will be reevaluated as 'existence probabilities' for the link itself. Given the alternative name, it shouldn't come as a surprise that they will now indicate the likelihood of a particular edge existing.

While this new interpretation may seem dramatically different compared to the previous realization, membership and existence are actually analogous to each other upon closer inspection. The overlap occurs since edges (in the IG) are **defined** to reflect how likely a high correlated distribution (**H**) generated the observations. In other words the presence of an edge will imply **H** as the generator (at high probability) and vice-versa[19]. Thus a high probability will indicate more confidence in the presence of the edge while a lower probability will indicate otherwise - a direct correspondence between ($p((x,y)$ from $\mathbf{H}$) and $p((x,y)$ from $\mathbf{L}$)) and (p(link exists) and p(link doesn't exists)).

With this new definition in place, the rest of the kernel discussion will be centered around this idea. This will ensure the applicability and validity of the resulting development.

3.6.1 Random Walk Base

SCRIP implements a type of 'walk kernel' to compute the inner product between two IGs on an infinite dimensional feature space. These kernels are characterized by transforming each featured graph into a 'walk count vector' before using a simple dot product to finalize the computation. Here the elements in the count vector correspond to the

[18] The probability that the observed gene pair comes from a high correlated distribution H.

[19] A lower probability will indicate that the edge doesn't exist and most likely isn't an observation from **H**.

number of times each unique labeled walk (from the graph) can be traversed. Since this set can be infinite in the case of cyclic and undirected graphs (i.e. traversing could never end), count vectors and their corresponding feature spaces can also be infinite dimensional.

However in order to explicitly construct the count vector, some walk kernels will select a few representative walks so that the dimensions are kept finite[11]. As a result the selection can simply limit walks past a certain length or more intelligently, leave out infrequent and/or tottering walks from the given graph[12]. On the other hand kernels that choose to tackle infinite dimensions will have to implement a weighting scheme corresponding to each walk since convergence becomes an issue otherwise (the summation as a result of taking the inner product between count vectors needs to converge).

In terms of SCRIP, the latter version of these walk kernels based on infinite dimensions was adopted. Specifically the setup resembles a 'random walk' function[13][20] where Markov probabilities[14][21] reflective of the aforementioned existence probabilities are used as the assigned weights. With that being said however, its implementation will be derived using the definition of a 'marginalized kernel[15]'. While this will initially differ from an inner product, these two concepts are actually analogous to each other as the final result can be re-expressed as a dot product weighted accordingly.

Nevertheless to introduce the derivation, the concept of labeled graphs and their reference to IGs will be required and hence are presented first.

3.6.2 Labeled Graphs

The IGs from SCRIP represent labeled graphs comprised of vertices and single edges (directed or undirected). If G is a labeled graph under this setup (with $|G|$ vertices) then:

- Let the vertices be **uniquely indexed** from 1 to $|G|$;
- Corresponding to these indices let $v(i)$ denote the **vertex label** for vertex index i and $e(i,j)$ denote the **edge label** for the edge between vertex index i and j.

---

[20] A random walk function is a type of walk kernel that defines the inner product between infinite dimensional count vectors. The uniqueness corresponds to the weighing scheme used.

[21] The Markov probabilities are in reference to each individual walk.

Figure 12 illustrates an example of the labeled graphs featured here. The distinction between vertex indices and labels can also be seen. Note that graphs with multiple edges between vertices and self loops are not considered.

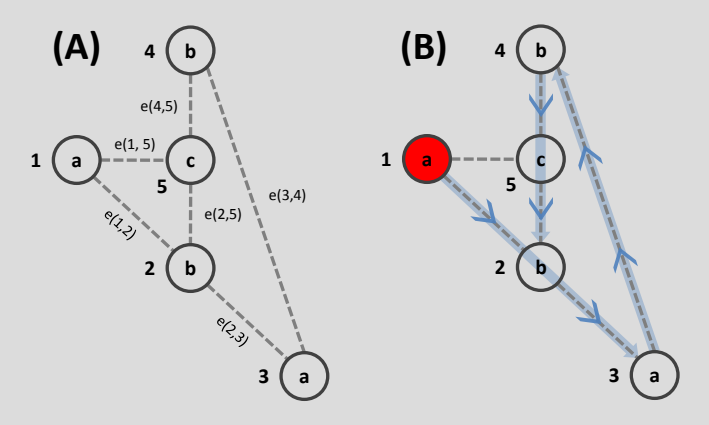


Figure 12: Labeled Graphs And Walks

(A) Assume graph G is given. Thus $|G| = 5$ and the vertex indices are $\{1,2,3,4,5\}$. Note that G has the following properties:

$$\text{Vertex Labels } \{v(1),v(2),v(3),v(4),v(5)\} = \{a,b,a,b,c\}$$

Since G is undirected, the pair of vertex indices indicated in each edge label can be reversed depending on the starting vertex. Thus:

$$\text{Edge label of } 1 \to 2 : e(1,2) = \text{Edge label of } 2 \to 1 : e(2,1)$$

(B) A walk variable (right side figure) is defined as a sequence of vertex indices such that there exists a link between any two consecutive indices. The walk variable w (starting at the red dot) with indices $\{1,2,3,4,5,2\}$ has the following properties:

$$\text{Labels of } w\text{: } \{v(1),e(1,2),v(2),e(2,3),v(3),e(3,4),v(4),e(4,5),v(5),e(5,2),v(2)\} = \{a,e(1,2),b,e(2,3),a,e(3,4),b,e(4,5),c,e(5,2),b\}$$

3.6.3 Marginalized Kernel To Walk Kernel

The marginalized kernel is a generalize setup that allows inner products to be constructed between non-vectoral instances (i.e. strings, graphs, trees, etc...). Due to its effortless extension to labeled graphs and consequently IGs, it was used to define a modified version of the random walk kernel. Not surprisingly the concept of 'random walks' or simply 'walk variables' will play an integral role throughout this entire discussion. Here they are defined as a sequence of vertex indices such that a link exists between any two consecutive entries. In Figure 12 for example, the walk variable W (starting at the red dot) has the following properties:

$$\begin{aligned}\text{Indices of } W\text{: } &\{1,2,3,4,5,2\}\\ \text{Labels of } W\text{: } &\{v(1),e(1,2),v(2),e(2,3),v(3),e(3,4),v(4),e(4,5),v(5),e(5,2),v(2)\}\\ =&\{a,e(1,2),b,e(2,3),a,e(3,4),b,e(4,5),c,e(5,2),b\}\end{aligned}$$

Consequently the marginalized kernel between IGs will assume the following (denote as G as the IG and W as the walk variable):

- The 'random component' is the walk variable W;
- The 'given component' is the graph variable G;
- The 'joint component' is joint variable $Z = (W, G)$.

The given names reflect what can or cannot be observed. For example while graphs are revealed[22], walks on the other hand are always hidden since the traversed paths are never known. Nevertheless with this setup, the formal definition of a marginalized kernel between two IGs $(G_1; G_2)$ and their corresponding set of all walks $(\mathscr{W}(G_1); \mathscr{W}(G_2))$[23] can be described as follows:

$$K(G_1,G_2) = \sum_{\mathscr{W}(G_1)} \sum_{\mathscr{W}(G_2)} p(W_1|G_1)p'(W_2|G_2)K_z(Z_1,Z_2) \tag{21}$$

Where K_z is the joint kernel referencing a particular walk from a particular graph.

In this formulation, the posterior probabilities $p(W_1|G_1)$ and $p'(W_2|G_2)$ represent the weights assigned to each instance of the joint kernel $K_z(Z_1,Z_2)$. In other words the marginalized kernel is simply a weighted expectation of K_z over all possible walks in G_1 and G_2. By choosing an appropriate representation of these posterior probabilities (or weights) used in conjunction to the IGs, a suitable kernel can be formulated. Ultimately this process starts with the selection of the weights before moving to the specification of the joint kernel. Once both pieces are obtained, Equation 21 (and the random walk kernel) will become fully specified.

3.6.4 Weight Specification

The posterior probability[24] of a walk $W = \{w_1, \ldots, w_l\}$ is constructed by recognizing it as a Markov chain. Since

---

[22]The graphs are revealed since they are physically obtainable through the merge method.

[23]$\mathbf{W}_1$ represents the set of all walks that can be traversed in G_1 while $\mathbf{W}_2$ for G_2 similarly.

[24]Posterior probability and weight are used interchangeably.

this assignment has a one to one correspondence with how W is generated, the process will be initially covered before presenting the associated weight. The procedure is as follows: (Assume $W = \{w_1, \ldots, w_l\}$ from graph G, where w_i refers to a sequence of natural numbers from 1 to $|G|$.)

- At the first step, w_1 is sampled from the initial probability distribution $p_s(w_1)$;
- Subsequently at the i^{th} step, the next vertex index w_i will be sampled subject to the transition probability $p_t(w_i|w_{i-1})$. Since the walk can also terminate with ending probability $p_q(w_i)$,

$$\sum_{j=1}^{|G|} p_t(j|w_{i-1}) + p_q(w_i) = 1$$

  Therefore at any given position, the walk must either transition or end.

Using the proposed initial, transition, and ending distributions, the posterior probability of W is described as:

$$p(W|G) = p_s(w_1) \cdot \prod_{i=2}^{l} \left\{ p_t(w_i|w_{i-1}) p_q(w_i) \right\} \tag{22}$$

Given no prior information, p_s and p_t can be set as uniform distributions over their respective feature spaces while p_q can be initialized as a constant. In the context of IGs however, SCRIP alters the transition and ending probabilities to accommodate the uncertainty associated with the presence of each edge. In particular this concept lends itself perfectly with the existence probabilities previously mentioned. These details are as follows:

- The ratio of transition probabilities from vertex index w_i to w_j and w_i to w_k should be proportional to the ratio of edge existence probabilities of w_i to w_j and w_i to w_k. In other words, if $e(w_i, w_j)$ has an existence probability 'x' times greater than the existence probability of $e(w_i, w_k)$, then the probability of transiting from w_i to w_j should also be 'x' times greater than transiting from w_i to w_k;
- The ending probability at vertex index w_i should reflect the uncertainty associated with the existence probabilities of all edges from w_i. In other words if edges from w_i has low existence probabilities, then the ending probability at vertex index w_i should be high since it is more likely those edges are truly missing (and hence nowhere to transition to). Using a similar argument if the existence probabilities are high, then the ending probability should be low (and hence a lot of vertices to transition to).

Using these guidelines the transition and ending probabilities at vertex index w^* were constructed as follows:

$$p_t(w_i|w^*) = \frac{p(e_i)}{k} \; \forall \; i \in 1, \ldots, k \tag{23}$$

$$p_q(w^*) = 1 - \sum_{i=1}^{k} p_t(w_i|w^*) \tag{24}$$

, assuming that:

- There were k potential edges $\{e_1, e_2, \ldots, e_k\}$ that connect w^* to k unique vertex indices $\{w_1, w_2, \ldots, w_k\}$ (which does not include w^*; no self loops);
- e_i: The edge label from w^* to w_i or $e(w^*, w_i)$;
- $p(e_i)$: The existence probability of $e(w^*, w_i)$;
- $p_t(w_i|w^*)$: The transition probability from w^* to w_i;
- $p_q(w^*)$ The ending probability at w^*.

In terms of the initial probability distribution, SCRIP will set it to the uniform distribution due to the lack of prior knowledge and biological explanation.

By plugging Equation 23 and 24 into Equation 22 this concludes the weight selection process and consequently integrates the existence probabilities and correlation design into the proposed workflow.

3.6.5 Joint Kernel Specification

With the weights in place, the final step is to define the joint kernel K_z referenced to in Equation 21. This will specify the binary relationship between walks from either graph through a comparison of their underlying joint variables. Under the same setup of the random, given, and joint components from Section 3.6.3 and the walk variable from Section 3.6.2, assume that walks W_1 and W_2 defined on IGs G_1 and G_2 have the following properties:

$$\begin{aligned}
&\text{Variable } W_1 = \{w_{11}, w_{12}, \ldots, w_{1l}\} \\
&\text{Variable } W_2 = \{w_{21}, w_{22}, \ldots, w_{2m}\} \\
&\quad \text{Label } W_1 = \{v(w_{11}), e(w_{11}, w_{12}), v(w_{12}), e(w_{12}, w_{13}), \ldots, e(w_{1(l-1)}, w_{1l}), v(w_{1l})\} \\
&\quad \text{Label } W_2 = \{v(w_{21}), e(w_{21}, w_{22}), v(w_{22}), e(w_{22}, w_{23}), \ldots, e(w_{2(m-1)}, w_{2m}), v(w_{2m})\}
\end{aligned}$$

Assuming this setup, the joint kernel will require two smaller kernel functions $K(v, v^*)$ and $K(e, e^*)$ to be defined on the vertex and edge labels respectively. Each label specific kernel will then compare the individual units within each given walk by assessing their similarity. For example if two vertex labels are identical at step x, then the label kernel should return a large numerical value in response to the concordance.

In the context of the IGs generated from the merge step, the labels associated with each walk will effectively correspond to the gene symbols and therefore carry no additional information within the string itself. For example, 'BRCA1' and 'BRCA2' are not 'more similar' simply because the two strings only differ by one character. Due to this

simple interpretation a straightforward comparison of the string labels is sufficient to construct the vertex and edge label kernels respectively. Specifically SCRIPT uses:

$$K(v,v^*) = I(v = v^*) \text{ and } K(e,e^*) = I(e = e^*) \tag{25}$$

, where $I(\ldots)$ represents the indicator function returning 1 if the argument holds and 0 otherwise. The joint kernel is then defined as the product of all labeled kernels corresponding to all individual components within each walk:

$$K_z(Z_1,Z_2) = K\left[v(w_{11}),v(w_{21})\right] \cdot \prod_{i=2}^{l} K\left[e(w_{1(i-1)},w_{1i}),e(w_{2(i-1)},w_{2i})\right] K\left[v(w_{1i}),v(w_{2i})\right] \tag{26}$$

, where $Z = (W,G)$. Note that the joint kernel will only return 1 when the labeled sequences from both walks matched up perfectly. In the event of mismatches or differential walk lengths ($l \neq m$), $K_z(Z_1,Z_2)$ will be defaulted to 0.

With this final result in Equation 26, the full specification of the joint kernel becomes completed.

3.6.6 Modified Random Walk Kernel

With the weight and joint kernel from Section 3.6.4 and refkernel5, the modified random walk kernel used in reference to SCRIP was obtained by plugging these results back into Equation 21. Thus the existing template becomes:

$$\begin{aligned}
K(G_1,G_2) &= \sum_{W_1 \in \mathscr{W}(G_1)} \sum_{W_2 \in \mathscr{W}(G_2)} p(W_1|G_1)p'(W_2|G_2)K_z(Z_1,Z_2) \\
&= \sum_{n=1} \sum_{W_1 \in \mathscr{W}_n(G_1)} \sum_{W_2 \in \mathscr{W}_n(G_2)} p(W_1|G_1)p'(W_2|G_2)K_z(Z_1,Z_2) \\
&= \sum_{n} \sum_{\mathbf{W}_1} \sum_{\mathbf{W}_2} \left\{ \left(p_s(w_{11}) \cdot \prod_{i=2}^{n} p_t(w_{1i}|w_{1(i-1)}) \cdot p_q(w_{1n}) \right) \times \left(p'_s(w_{21}) \cdot \prod_{i=2}^{n} p'_t(w_{2i}|w_{2(i-1)}) \cdot p'_q(w_{2n}) \right) \times \right. \\
&\qquad \left. \left(K\left[v(w_{11}),v(w_{21})\right] \cdot \prod_{i=1}^{n} K\left[e(w_{1(k-1)},w_{1k}),e(w_{2(k-1)},w_{2k})\right] K\left[v(w_{1k}),v(w_{2k})\right] \right) \right\}
\end{aligned} \tag{27}$$

, where: $\mathscr{W}(G)$ represents all walks in graph G;

$\mathscr{W}_n(G)$ represents all length n walks in graph G;

$\sum_{W \in \mathscr{W}_n(G)} := \sum_{w_1=1}^{|G|} \cdots \sum_{w_n=1}^{|G|}$.

The proposed random walk kernel will closely resemble a weighted expectation of K_z over all possible walks in graphs G_1 and G_2. While this effectively summarizes the computation, straightforward enumeration of Equation 27 is not feasible since n spans from 1 to infinity. Fortunately by taking advantage of the 'product graph[16]' $G_1 \times G_2$ induced by the tensor product[17] (between graphs G_1 and G_2), a solution to Equation 27 that can furthermore be computed in polynomial time becomes derivable. Thus assuming that $G_1 = (V_1,E_1)$ and $G_2 = (V_2,E_2)$ are two graphs denoted by their vertex (V_1,V_2) and edge (E_1,E_2) labels respectively, the product graph $G_1 \times G_2$ is defined as the graph $G = (V,E)$ with:

$$V = \{(v_1, v_2) \in V_1 \times V_2 : v_1 \text{ and } v_2 \text{ have the same label}\}$$
$$E = \{((v_1, v_2), (v_1', v_2')) \in V \times V : (v_1, v_1') \in E_1 \text{ and } (v_2, v_2') \in E_2\}$$

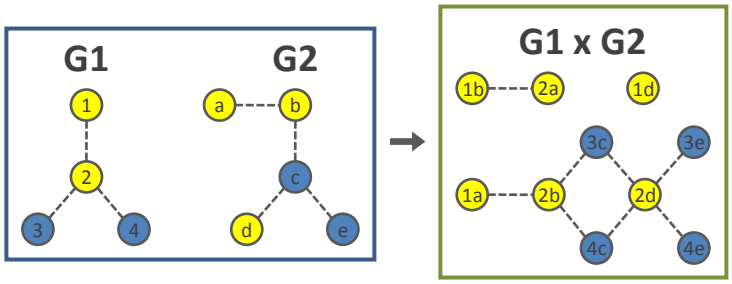


Figure 13: Product Graph Example

The product graph between G1 and G2 is depicted in G1×G2. Note that all walks shared between G1 and G2 could be found in G1×G2 and vice-versa.

The unique property of the product graph that can subsequently assist Equation 27 is the bijection claiming:

1. Pairs of walks $W_1 \in \mathscr{W}(G_1)$ and $W_2 \in \mathscr{W}(G_2)$ sharing same labels,
2. Walks on the product graph: $W \in \mathscr{W}(G_1 \times G_2)$.

Therefore all shared walks between the two IGs are in the product graph and all walks in the product graph are also shared between the two IGs. This property will have resounding consequences since the walk kernel in Equation 27 can then be simplified to the following structure: (Note: walks in the product graph $W \in \mathscr{W}_n(G_1 \times G_2) = (w_1, \ldots, w_n)$)

$$\begin{aligned}
K(G_1, G_2) &= \sum_n \sum_{\mathbf{W}_1} \sum_{\mathbf{W}_2} \Bigg\{ p_s(w_{11}) p_s'(w_{21}) K\left[v(w_{11}), v(w_{21})\right] \times p_q(w_{1n}) p_q'(w_{2n}) \times \\
&\qquad \prod_{i=2}^{n} p_t(w_{1i}|w_{1(i-1)}) p_t'(w_{2i}|w_{2(i-1)}) K\left[e(w_{1(i-1)}, w_{1i}), e(w_{2(i-1)}, w_{2i})\right] K\left[v(w_{1i}), v(w_{2i})\right] \Bigg\} \\
&= \sum_n \sum_{\mathbf{W}_1} \sum_{\mathbf{W}_2} \left\{ s(w_{11}, w_{21}) \cdot q(w_{1n}, w_{2n}) \cdot \prod_{i=2}^{n} t(w_{1i}, w_{1(i-1)}, w_{2i}, w_{2(i-1)}) \right\} \\
&= \sum_n \sum_{W \in \mathscr{W}_n(G_1 \times G_2)} \left\{ s(w_1) \cdot q(w_n) \cdot \prod_{i=2}^{n} t(w_i, w_{i-1}) \right\} \qquad (28)
\end{aligned}$$

, where:

$$\begin{aligned}
s(w_{11}, w_{21}) &= p_s(w_{11}) p_s'(w_{21}) K\left[v(w_{11}), v(w_{21})\right] \\
s(w_1) &= p_s(w_1) p_s'(w_1) \\
q(w_{1n}, w_{2n}) &= p_q(w_{1n}) p_q'(w_{2n}) \\
q(w_n) &= p_q(w_n) p_q'(w_n) \\
t(w_{1i}, w_{1(i-1)}, w_{2i}, w_{2(i-1)}) &= p_t(w_{1i}|w_{1(i-1)}) p_t'(w_{2i}|w_{2(i-1)}) \times \\
&\qquad K\left[e(w_{1(i-1)}, w_{1i}), e(w_{2(i-1)}, w_{2i})\right] K\left[v(w_{1i}), v(w_{2i})\right] \\
t(w_i, w_{i-1}) &= p_t(w_i|w_{i-1}) p_t'(w_i|w_{i-1})
\end{aligned}$$

Since the joint kernel K_z is defined as a series of indicator functions back in Section 3.6.5, $s(w_{11}, w_{21}) \neq 0$ and $\prod_{i=1}^{n} t(w_{1i}, w_{1(i-1)}, w_{2i}, w_{2(i-1)}) \neq 0$ if and only the labels corresponding to the compared walks (W_1 and W_2) match up perfectly. Hence the final equality in Equation 28 can be established since the summation only needs to run through the set of shared walks. Consequently the summation space reduces to the walks exclusive to the product graph by definition. By rearranging the terms around, the random walk kernel has the following final structure:

$$\begin{aligned} K(G_1, G_2) &= \sum_{n} \sum_{w \in \mathscr{W}_n(G_1 \times G_2)} \left\{ s(w_1) \cdot q(w_n) \cdot \prod_{i=2}^{n} t(w_i, w_{i-1}) \right\} \\ &= \sum_{n=1}^{\infty} (\Lambda_s \circ \Lambda_s')^T (\Lambda_t \circ \Lambda_t')^n (\Lambda_q \circ \Lambda_q') \\ &= (\Lambda_s \circ \Lambda_s')^T \left((\mathbf{I} - \Lambda_t \circ \Lambda_t')^{-1} - \mathbf{I} \right) (\Lambda_q \circ \Lambda_q') \end{aligned} \tag{29}$$

, where: Λ_s is the vector of starting probabilities with i^{th} element $p_s(i)$ (in G_1)

Λ_q is the vector of ending probabilities with i^{th} element $p_q(i)$ (in G_1)

Λ_t is the matrix of transition probabilities with i^{th} row and j^{th} column $p_t(i|j)$ (in G_1)

$\Lambda_s', \Lambda_q',$ and Λ_t' carry the same meaning but in G_2

Thus by taking advantage of the product graph, a closed form solution to the proposed random walk kernel was obtained. With the only requirement being matrix inversion, it becomes a relatively simple procedure in comparison to the infinite summation initially introduced in Equation 27.

3.6.7 Correspondences Between Marginalized Kernel And Dot Product

The proposed random walk kernel formulated in Section 3.6.6 was constructed from the template of a marginalized kernel. While this may seem different compared to a dot product between 'count vectors[25]', the two ideas are actually analogues of each other as mentioned in Section 3.6.1. To show this correspondence, the previously formulated random walk kernel in Equation 29 will be re-expressed as a dot product between infinite dimensional count vectors weighted according to Markov posterior probabilities. Note that the derivation here will be presented in reverse order. Thus starting from the dot product, Equation 29 will be obtained by working backwards. This process starts with the count vector $\phi(G)$ of graph G:

$$\phi(G) = [\phi_s(G)]_{s \in S(G)}$$

[25]The vectors are vectoral representations of each IG.

, where: $S(G)$ represents the set of all walk labels (in comparison to $\mathscr{W}(G)$ which represents the set of all indexed walks);

$\phi_s(G) = \sum_{W \in \mathscr{W}(G)} \lambda_G(W) \cdot 1[s \text{ is labeled walk of } w]$;

$\lambda_G(W)$ represents the weight associated with a walk variable (sequence of walk indices) $W \in \mathscr{W}(G)$.

In this setup each element of $\phi(G)$ corresponds to the number of times a given labeled walk can be traversed in the graph (different sequences of walk indices may end up having the same walk label). The count will then be weighted by the summation of all weights $\lambda_G(W)$ that indexed this particular walk label. Thus following the given notation the dot product between weighted count vectors resembling G_1 and G_2 can be described as:

$$\begin{aligned}
K_{\text{walk}}(G_1, G_2) &= \sum_{s \in S(G)} \phi_s(G_1) \cdot \phi_s(G_2) \\
&= \sum_{s \in S(G)} \left[\left(\sum_{W_1 \in \mathscr{W}(G_1)} \lambda_{G_1}(W_1)[s = \text{label of } W_1] \right) \left(\sum_{W_2 \in \mathscr{W}(G_2)} \lambda_{G_2}(W_2)[s = \text{label of } W_2] \right) \right] \\
&= \sum_{W_1 \in \mathscr{W}(G_1)} \sum_{W_2 \in \mathscr{W}(G_2)} \lambda_{G_1}(W_1) \cdot \lambda_{G_2}(W_2) \cdot [\text{label of } W_1 = \text{label of } W_2] \\
&= \sum_{W_1 \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} \sum_{W_2 \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} \lambda_{G_1}(W_1) \cdot \lambda_{G_2}(W_2) \\
&= \sum_{W \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} c_W \cdot \lambda_{G_1}(W) \cdot \lambda_{G_2}(W) \\
&= \sum_{W \in \mathscr{W}(G_1 \times G_2)} \lambda_{G_1 \times G_2}(W) \\
&= \sum_{n=1}^{\infty} \sum_{W \in \mathscr{W}_n(G_1 \times G_2)} \lambda_{G_1 \times G_2}(W)
\end{aligned} \tag{30}$$

Thus according to this final form, the dot product simplifies to the summation of all weights associated with walks in the product graph. By subsequently setting these weights as the multiplication between Markov walk probabilities in graphs G_1 and G_2, the result will then resemble the random walk kernel presented in Section 3.6.6. Hence the correspondence between an inner product and the Marginalized kernel is established.

References

[1] L. Van't Veer, R. Bernards, *Nature* **452**, 564 (2008).

[2] D. Slonim, *Nature genetics* **32**, 502 (2002).

[3] K. Parthasarathy, K. Schmidt, *Positive definite kernels, continuous tensor products, and central limit theorems of probability theory* (Springer-Verlag, 1972).

[4] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[5] S. Pan, K. So, J. Rahmeh, *ACM SIGPLAN Notices* (ACM, 1992), vol. 27, pp. 76–84.

[6] T. Gärtner, P. Flach, S. Wrobel, *Learning Theory and Kernel Machines* pp. 129–143 (2003).

[7] J. Ramon, T. Gärtner, *First International Workshop on Mining Graphs, Trees and Sequences* (2003), pp. 65–74.

[8] T. Golub, *et al.*, *science* **286**, 531 (1999).

[9] G. Parmigiani, E. Garett, R. Irizarry, S. Zeger, *The analysis of gene expression data: methods and software* (Springer, 2003).

[10] I. Dhillon, Y. Guan, B. Kulis, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2004), pp. 551–556.

[11] A. Smola, R. Kondor, *Learning theory and kernel machines* pp. 144–158 (2003).

[12] P. Mahé, N. Ueda, T. Akutsu, J. Perret, J. Vert, *Journal of Chemical Information and Modeling* **45**, 939 (2005).

[13] K. Borgwardt, *et al.*, *Bioinformatics* **21**, i47 (2005).

[14] K. Chung, *et al.*, *Markov Chains with Stationary Transition* (New York: Springer, 1960).

[15] H. Kashima, K. Tsuda, A. Inokuchi, *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* (2003), vol. 20, p. 321.

[16] T. Feder, *Journal of graph theory* **16**, 467 (1992).

[17] M. Marcus, B. Moyls, *Pacific Journal of Mathematics* **9**, 1215 (1959).

Kernel Simulation

Chapter 4

This chapter details a simulation designed to investigate (modified random walk) kernel competence. Under the working assumptions of '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP), the proposed kernel from Chapter 3 was used to classify simulated graphs in an attempt to elucidate data-to-kernel based interactions. This is necessary in order to validate of the kernel used in reference to SCRIP.

4.1 Introduction

'**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) was developed as a hybrid tool between mean- and correlation-based classification. The corresponding workflow is defined by separate kernel functions with respect to SVM[1] and its optimization theory. In particular the RBF[2] and random walk functions[3] were implemented as inner products in reference to the mean and correlation signatures. Each one effectively designates the metric in which subject similarity is evaluated within their respective setups.

Due to the heavy implications of these kernels, their efficacy becomes an integral part to the success of SCRIP. While this quality is guaranteed from the RBF kernel, the walk kernel in comparison lacks these technicalities due to its unpolished nature as a strict theoretical development. Therefore in response to this concern, the following simulations were proposed in order to elucidate the kernel's ability to interact with SVM, the IGs, and SCRIP in general.

To design such simulation, the emphasis was subsequently placed on replicating a real-life application so the results can be generalized as future references. Consequently the expression vectors were designated as the simulated objects since the data often represents an intuitive starting point. Therefore by way of various parameters that jointly define the co-expression signatures and noise level, pseudo GEPs were generated (through a multivariate normal distribution) at the beginning of each simulation round. To then evaluate kernel performance under the given parameter settings, the correlation signature was fitted to these points and the prediction accuracy was used as a referencing summary. Not surprisingly these accuracies will end up forming a comparative base across all simulations thereby pinpointing the most ideal conditions for this particular classification tool.

The following chapter details the proposed simulations intended to verify kernel efficacy in response to the correlation signature. With that being said, the following discussion starts with the simulation design. Afterwards the final results and their interpretation will follow to effectively conclude this section.

4.2 Simulation Setup

Under the goal of simulating expression vectors as a means to test the walk kernel, a multivariate normal distribution (MVN) was adopted due to the assumed normality of each gene and consequently the need to model correlation-based signatures. Therefore by using a set of parameters including the dimension size (number of simulated genes),

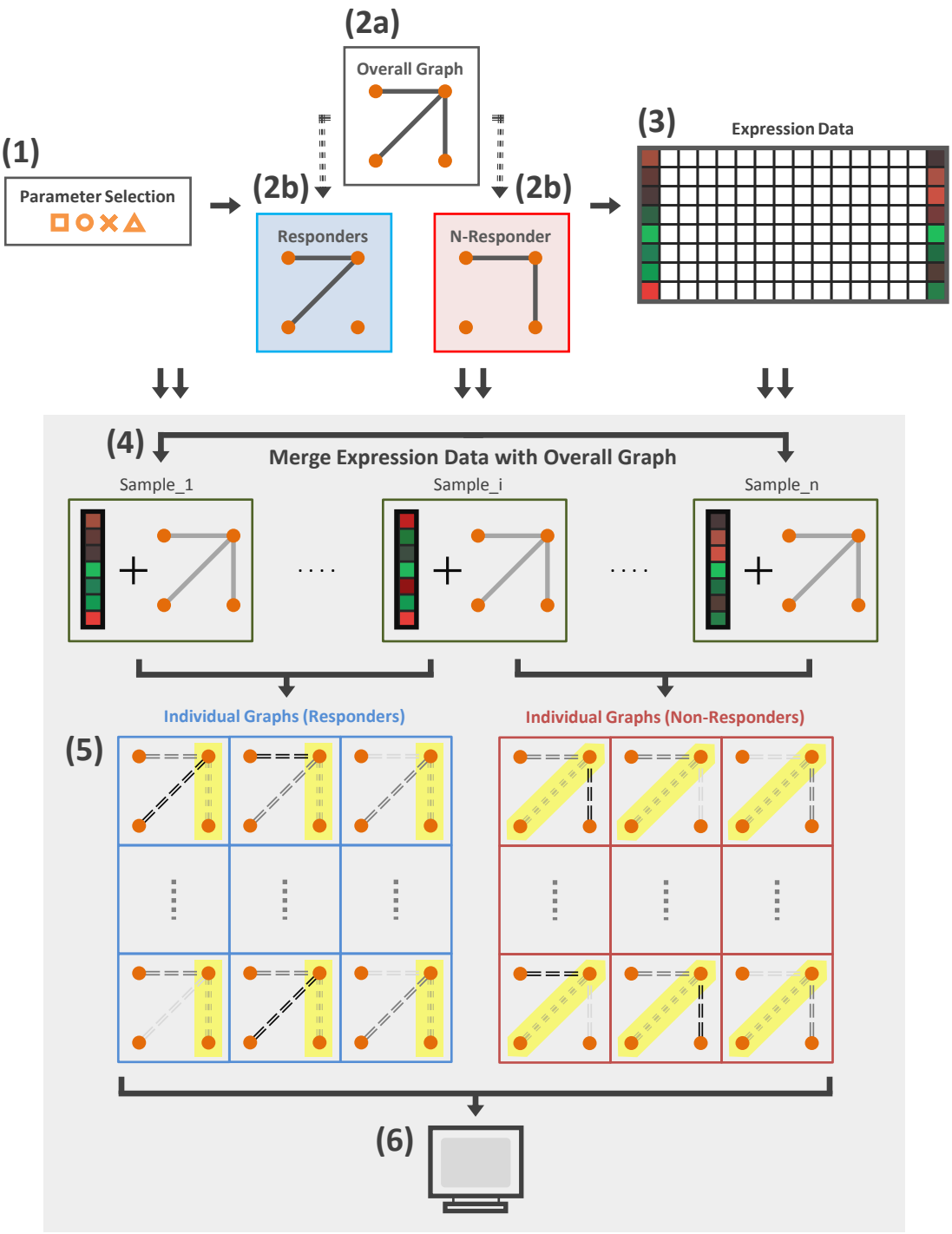


Figure 14: Simulation Workflow

After designating a set of parameters (1) which would consequently structure the overall (2a) and response specific graphs (2b), pseudo gene expression vectors (3) were simulated through the help of a MVN. Once these vectors were obtained, the correlation signature of SCRIP (4-6) was applied (in the same fashion as applications) to evaluate kernel efficacy under the given parameter settings. A description of steps 4, 5, and 6, could be seen in Figure 8.

mean/variance vectors, and correlation matrix, MVN observations were generated to form a hypothetical dataset for the analysis. In particular the only parameters altered across simulations were the dimension size and correlation matrices since they can effectively induce all noise and signal levels intended for investigation. The mean and variance vectors on the other hand were set to 0 and 1 respectively since all real-life applications will be based on the standardized values anyways. Thus instead of simulating the complexity involved with raw expression values, their standardized forms were instead used as a attempt to simplify the entire process.

For a given parameter setting, the simulation followed:

- An overall graph (G) was created with various nodes and a random placement of edges amongst them (20% to 25% of all possible edges);
- From G, two response specific graphs (RSGs) were formed to contrast at a fraction of those edges. Increasing this fraction therefore magnifies the difference between the RSGs and consequently the implied signal level;
- The adjacency matrix representations of these RSGs were then transformed into valid correlation matrices using spectral decomposition[4]. Thus each class was assigned a personalized correlation signature;
- With the proposed mean, variance, and correlation matrices, standardized expression vectors corresponding to each class were generated;
- The correlation signature of SCRIP was applied to these vectors using G as the pseudo pathway. Additionally the mean signature was also applied to form a comparative base. In both cases the prediction accuracies were recorded to signal the end of the given simulation round.

Thus by repeating the aforementioned outline across a wide spectrum of parameter combinations, kernel efficacy was evaluated by simply assessing the returned accuracies. As a result questions regarding the kernel's base efficiency and how it responds to varying degrees of noise and signal were all elucidated in response to the correlation signature of SCRIP.

In the remainder of this chapter, the details that map out the proposed outline will be presented.

4.2.1 Overall Graph $\rightarrow$ Correlation Matrix

The overall graph (OG) used in reference to any simulation was initialized as an $n \times n$ adjacency matrix G with

a total of l randomly placed, undirected edges. Under their construct these edges highlighted potential class-specific discrepancies between the correlation statuses (high vs. low) of each referenced pair. Since this exactly defines the combined information across both classes, the corresponding response specific graphs (RSGs) were created by randomly perturbing a proportion (p) of these edges so that the results only differed at those exact positions.

To use the resulting RSGs as correlation matrices (in the context of the MVN), modifications were subsequently required to address their adjacency-based forms. Fortunately since the edges within these graphs already designate the presence/absence of correlation-based statuses (presence implies $\rho = 0.85$; absence implies $\rho = 0.05$), minor adjustments were therefore only required to finalize this procedure. Specifically they were carried out as follows:

- All entries in the adjacency matrix with value 1 (or the presence of an edge) were changed to 0.85 (indicating the presence of a high correlation between the two features);
- All entries in the adjacency matrix with value 0 (or the absence of an edge) were changed to 0.05 (indicating the presence of a low correlation between the two features);
- All entries in the diagonal were changed to 1.

Since these altered adjacency matrices must also conform to a positive semi-definite criterion, final adjustments[5] were therefore necessary in order to complete its validation process[26]. Specifically the goal here was to pinpoint a legitimate alternative that furthermore exhibits the most affinity to its original form. To do so, spectral decomposition was employed and carried out as follows (for argument sake, suppose A represents the altered adjacency matrix):

- The eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ and eigenvectors $\{\mathbf{s}_1, \ldots, \mathbf{s}_p\}$ corresponding to A were calculated;
- The diagonal eigenvalue matrix $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$ was created;
- The eigenvector matrix $\mathbf{S} = \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{bmatrix}$ was created;
- Each negative element in $\mathbf{\Lambda}$ was replaced with 0 and the new diagonal matrix was denoted as $\hat{\mathbf{\Lambda}}$. Thus the new eigenvalues $\{\hat{\lambda_1}, \ldots, \hat{\lambda_n}\}$ had the following non-negative property: $\hat{\lambda}_i \geq 0 \; \forall \; i \in \{1, \ldots, n\}$;
- The diagonal scaling matrix $\mathbf{T}$ was calculated where $t_{ii} = \left(\sum_{k=1}^{n} s_{ik}^2 \hat{\lambda}_k \right)^{-1} \; \forall \; i \in \{1, \ldots, n\}$;

[26] On most occasions these altered adjacency matrices don't conform to a positive semi-definite criterion.

- The validated correlation matrix $\hat{A}$ was finally constructed as follows: $\hat{A} = \left[\mathbf{t}\cdot\mathbf{S}\cdot\hat{\boldsymbol{\lambda}}\right]\left[\mathbf{t}\cdot\mathbf{S}\cdot\hat{\boldsymbol{\lambda}}\right]^T$, where $\mathbf{t} = \mathbf{T}^{1/2}$ and $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\Lambda}}^{1/2}$.

Thus by apply the transformation to both RSGs, the reconstruction of their correlation matrices was finalized. Consequently they were ready for simulation use with respect to the parameters initializing their underlying OG.

4.3 Parameter Selection

The interactions between correlation structure, signal-to-noise ratio, and sample size were investigated in this simulation. To subsequently generate the expression profiles corresponding to these settings, four parameters were used in conjunction to the MVN. Here the use of multiple inputs enabled a finer degree of tuning amongst the variables of interest. In particular the parameters include:

- N: The sample size of each simulated class (thus a total of $2N$ samples were generated per simulation);
- n: The number of OG nodes;
- l: The number of OG links;
- p: The proportion of OG links that differ between the RSGs.

Interpretation wise, l and p defined the correlation matrices used in reference to each class and consequently the intended levels of signal and noise between both sets of vectors. For example by coupling a larger l with a large p, the intended signal level becomes amplified since an excess of differential edges would appear between the two classes. As a result the corresponding RSGs would gain more individuality and correspondingly, yield a greater distinction (based on correlation) between their simulated vectors. Through a similar argument, a small l together with a small p would then induce excess noise. On the other hand N and n were used to define the simulation size; and while incapable of directly specifying the signal level, acted instead as a promoter (or minimizer) in response to a given l and p combination. Thus under these four parameters, any simulation setup could be described and consequently carried out.

Nevertheless while the combination of N, n, l, and p enabled access to an entire spectrum of simulation designs, their corresponding domain was unfortunately too extensive thus making a thorough exploration an impractical proposition. To thereby balance computation practicality with ambition, a compromise was settled by reserving simulation to the most realistic designs amongst the greater list. Consequently this simplified the initial parameter combinations down to two generalized sets. First, N and n were set to 100 and 10 respectively while l and p varied. Here the effects

of varying degrees of signal within a small pathway were elucidated using a reasonably sized dataset. And second, with an exchange of roles, l and p were fixed under different values of N and n. This consequently provided insight to the size of the study and its interaction with a common OG/RSG setup[27].

Thus in summary:

- First simulation $\rightarrow N = 100$, $n = 10$,
$$l = \{0.2, \ldots, 0.4\} \times \tfrac{1}{2}n(n-1), \text{ and } p = \{0.11, \ldots, 0.2\} \times l;$$
- Second simulation $\rightarrow l = 0.27 \times \tfrac{1}{2}n(n-1)$, $p = 0.2 \times l$,
$$N = \{100,\ 200,\ 300\},\ n = \{10,\ 20,\ 30,\ 40,\ 50\}$$

4.4 Simulation Count And Split

With the specification of these parameters completed, the simulation was then replicated 99 times on each individual combination where 70% and 30% of the data was used for training and testing respectively. By using their averaged test results as the referencing summary, robustness was guaranteed in conjunction to these returned values.

4.5 Results And Discussion

Under various parameter combinations highlighting correlation structure, signal-to-noise ratio, and sample size, the aforementioned simulation process was carried out in order to elucidate inherent technicalities of the derived kernel. Here the primary goals were to: (1) Evaluating kernel performance across a multitude of data settings; and (2) Compare these results with the radial basis function (RBF) to verify efficacy. As it pertains to both tasks, the returned accuracies from each simulation served as an effective evaluation criterion.

Generally speaking the walk kernel outperformed the RBF across all parameter combinations as expected. As a testament to the kernel's performance, the correlation signature returned an averaged prediction accuracy of 0.705

[27]Common refers to the size of the OG or the number of nodes and links.

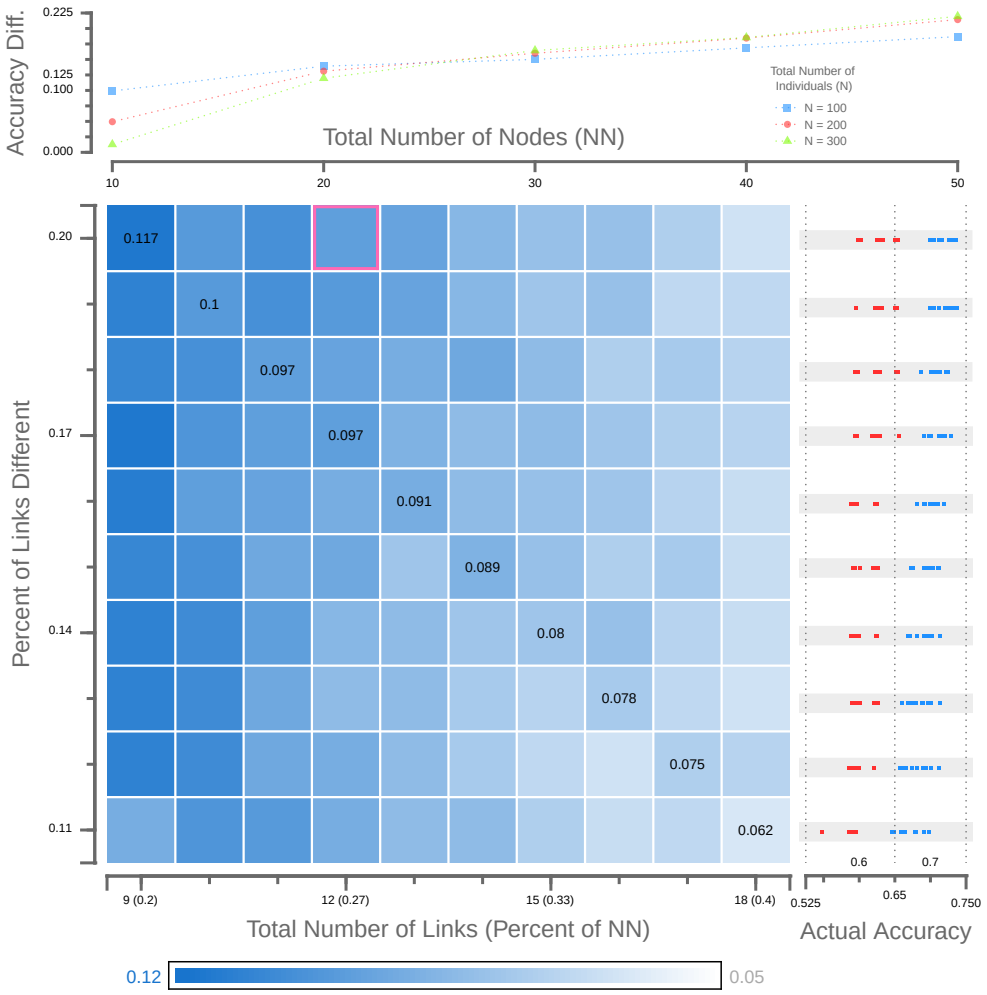


Figure 15: Visualization Of Simulation Results

The lower right heatmap displays the difference in prediction accuracy (PA) between the correlation signature of SCRIP and standard SVM ($N = 100$; $n = 10$; l and p vary). In each combination of l and p, SCRIP outperforms the aforementioned approach. The values on the horizontal tiles indicate the numerical difference in PAs and the bottom right panel indicates the actual PAs for each combination of LD (blue dots: correlation based; red dots: mean based). The top scatterpot displays the difference in PA between SCRIP and standard SVM ($l = 27\%$; $p = 20\%$, N and n vary). The highlighted blue tile (in pink) from the heatmap indicates the combination of l and p used in all simulations for the scatterplot.

N: Number of simulated vectors per response group;
n: Number of simulated nodes for the OG;
l: Number of simulated links in the OG (percentage of the maximum number of links);
p: Percentage of l that vary between each RSG.

(median 0.701) far exceeding its mean counterpart at 0.609 (median 0.601). In fact under the most favorable conditions for the walk kernel ($N = 300$, $n = 50$, $l = 62$, $p = 0.2$), the accuracy difference would actually widen to 0.231 thus telling a compelling story for the kernel's efficacy in light of the proposed assumptions (0.797 vs. 0.567). Thus these results conclusively state the derivation's capability to process a genuine correlation signal for downstream analysis in SVM.

With kernel efficacy established, the question then turns to its cross sectional performance as it pertains to the collection of simulated scenarios. Since this was in part elucidated under the setup of two generalized parameter sets (as mentioned in Section 4.3), it will therefore be presented here on an individual basis. Thus:

- Under the scenario where simulation size was predefined ($N = 100$; $n = 10$), kernel performance was optimized under the concurrent decrease of links and increase of signal both in reference to the OG (l and p respectively). From a computation perspective this makes sense since fewer links would inevitably highlight the presence of the remainders, which if just so happens to be richer in signal, would consequently benefit the classification process. As a result this states the merit for sparse networks under the assumption that they contain a high proportion of informative links. On the other hand any deviation from this given specification (i.e. increase in link or decrease in signal) would only serve to undermine the kernel's performance.

  Note that under this parameter set, the maximum and minimum prediction accuracies were achieved at 0.711 ($l = 9$; $p = 0.2$) and 0.649 ($l = 18$; $p = 0.1$) respectively.

- Under the scenario where the signal level was predefined (OG set; $l = 0.27 \cdot n(n-1)$; $p = 0.2\%$), kernel performance was evaluated through the interactions between (OG) dimension and sample size (N and n respectively). Here the results were dissected from two separate angles (each parameter was assessed assuming consistency from the other) so that their interpretations could be presented in a logical fashion. First as the dimensions shrank, kernel performance consequently faltered amongst simulations using the same sample size (hold N; move n). And second, in a reversal of roles, as sample sizes decrease across identical OGs, performance ironically improved at first (for smaller OGs) before faltering according to intuition later on (for the larger OGs). While alarming, the initial case could probably be attributed to the randomness resulting from a scarcity of input nodes and edges.

  Note that under this parameter set, the maximum and minimum prediction accuracies were achieved at 0.797 ($N = 300$; $n = 50$) and 0.709 ($N = 100$; $n = 20$) respectively.

As a consequence of these findings, the ideal working conditions for the kernel were extrapolated by a simple

cross referencing procedure. Consequently the full story behind parameter fluctuations and their repercussions were obtained and described as follows:

- Assuming that a large, sparse, and highly informative pathway (n large, l small, p large) could be analyzed using a sufficient sample size, kernel performance should be optimized with a prediction accuracy close to 0.85 (and difference with the mean signature close to 0.3).

- On the flip side, assuming that a small, dense, and non-informative pathway (n small, l large, p small) was furthermore coupled with an insufficient sample size, kernel performance would then resemble nothing more than a random guess with accuracy around 0.5.

Since it is perceivable that any application setting will lie somewhere between complete junk and the best case scenario, the simulation makes a compelling case for the correlation signature assuming the validity of the proposed assumptions.

References

[1] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[2] I. Dhillon, Y. Guan, B. Kulis, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2004), pp. 551–556.

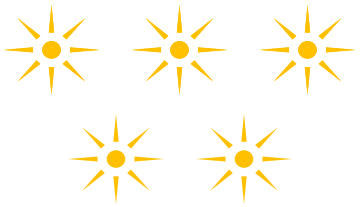
[3] K. Borgwardt, *et al.*, *Bioinformatics* **21**, i47 (2005).

[4] G. Partyka, J. Gridley, J. Lopez, *Amoco Production Company, Leading Edge* **18**, 353 (1999).

[5] R. Rebonato, P. Jäckel, *Available at SSRN* (2011).

Response Prediction Application

Chapter 5



This chapter details the application of '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) to a response prediction problem involving multiple myeloma patients. After introducing the dataset, the discussion shifts to the modeling process. First, the mean signature of SCRIP will be detailed followed by the correlation counterpart. Upon completion, the final predictions will then be presented by combining both pieces of inference.

5.1 Introduction

With the methodological development of '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) completed, the emphasis will now be shifted in the direction of its application counterpart. Here multiple myeloma[1] (MM) patients from the HOVON-65/GMMG-HD4 trial[2] were analyzed using the developed framework. As a result a mean- and correlation-based response signature was trained under the premise of gene expression data.

With this choice of application, SCRIP also receives its first test in light of the associated challenges. Most notably its use of expression data exposes a routine problem concerning the inadequacy of these profiles to model MM-based response: Capped accuracy at 70% regardless of treatment type[3]. Nevertheless while such restriction will seemingly handicap any analysis, this impediment ironically serves as a definitive evaluation of SCRIP right off the bat. Because if the proposed method works as intended, then the corresponding prediction accuracy should exceed the given benchmark as a simple confirmation of its relevance and potential.

In terms of the analysis, the following chapter details the interventions and justification behind the training process. This not only provides intuition for the current analysis, but also ensures effortless extension of future applications. In preparation for the remainder of this discussion however, the dataset used in conjunction to the application will be first presented. This entails the data split, array preprocessing, and feature filtering. Afterwards the modeling and final predictions will follow to conclude the application part of this thesis.

5.2 Dataset Characteristics

The application dataset was based on the gene expression profiling (GEP) of 320 newly diagnosed MM patients from the Dutch-Belgian/German HOVON-65/GMMG-HD4 clinical trials. Here all subjects were free of any prior treatment and were uniformly assigned to a three-drug regimen (VAD/PAD[4]) followed by autologous stem cell transplantation[5] (ASCT). Since the response information (response label) was only available on 282 of the 320 subjects, the dataset used in conjunction to this analysis therefore only represented a subset from the initial cohort.

Here the expression data was obtained from the Gene Expression Omnibus[6] through accession number GSE19784. The response information was however directly obtained from the authors since it remains yet unpublished.

5.2.1 Preprocessing

GEP from the 282 samples were hybridized to the Affymetrix Gene Chip Human Genome U133 plus 2.0 arrays. To subsequently arrive at the signal intensities, quality control, preprocessing, and normalization were done in separate software packages. First, quality control was done using the GeneChip Operating Software. Here a combination between 'scaling factor' and 'percentage of genes present' was used to remove arrays with a scaling factoring difference greater than 3 or a gene presence less than 20%. Afterwards preprocessing and normalization was carried out using the GCRMA[7] package in Partek Genomics Suite version 6.4[8]. Upon completion a secondary quality control was carried out using the AffyPLM[9] package from Bioconductor[10]. Here the relative log expression and normalized unscaled standard errors (NUSE) jointly removed arrays exhibiting a NUSE value greater than 1.05 to go along with an aberrant expression plot.

5.2.2 Feature Filtering

All $54,675$ probes in the preprocessed dataset ($n = 282$) were filtered for quality control before any analysis took place. To remove extremities and non-expressed features, each probe was subjected to the following selection criterion:

- For each probe the 90^{th} percentile was calculated across all 282 subjects;
- Of the ordered percentiles, only the probes falling in the top 50% were selected for analysis.

Note that '50%' was designated in order to reflect the percentage of non-expressed genes from any given tissue type[11] (roughly half in total). This resulted in the final selection of $26,616$ probes.

5.2.3 Dataset Split

All 282 subjects were randomly partitioned into three nonintersecting groups for training purposes. 60% of the subjects (169 subjects) formed the first group while 20% formed each of the later two (57 and 56 subjects respectively). All calculations were carried out in R[12].

5.3 Class Imbalance

With any analysis pertaining to binary classification, 'class imbalance' poses a practical difficulty across most applications. In effect as a particular class becomes increasingly outnumbered (usually the more interesting of the two), learning usually succumbs to a common methodological design that will inevitably emphasize the prevalence of a particular signal[13–17]. Thus unless highly unusual conditions apply (i.e. The data is already perfectly separable), the rare events will become overwhelmed and subsequently be ignored in any analysis, SCRIP included.

Unfortunately the HOVON dataset also falters to the aforementioned issue; though not entirely a surprise due to the prevalence of such problem. Fortunately however, the scarcity of responders (76 out of 282 subjects) in this case didn't end up jeopardizing the entire analysis in part due to the flexibility built into SCRIP. Specifically since the associated training isn't fixated on a set of rigid procedures, any custom built strategy designed to maximize the potential of a given dataset can therefore be adopted. Consequently the prediction here was predicated on 'favoring the responders[28]'.

This proposed training alternative essentially highlights the rare cases at the expense of undertaking a reasonable amount of false positives as a compensation for their inconsistency. To control the inevitable inflation, the strategy took advantage of SCRIP's setup and was only applied to each individual signature (mean and correlation model) before a union of their results acted as the quality control. While not capable of filtering away all superimposed false positives, the trade off between the deliberate favoritism and error rate still proved to be advantageous and was therefore adopted.

5.4 Mean Signature Application Process

The mean signature used in reference to SCRIP was an adaption taken from the classical modeling approach[18]. In an attempt to thoroughly explore mean-based classification, the signature was comprised of two mean-based models separately trained in SVM. The first model used log values[29] while the second incorporated standardization[30] into the

[28]Conduct the training so that more response predictions are obtained.

[29]Log base 2.

[30]Standardization was preformed across the samples for each input feature or gene.

mix. In both setups the RBF kernel[19] defined the inner product between vectoral representations of each subject[31]. Upon completion the union of their results then formalized the mean-based signature of SCRIP.

The following section outlines the application process. This includes the data split, feature filtering, model selection, and result combination. The reason for proposing two separate models will also be brought up throughout this discussion.

5.4.1 Data Split

In the same fashion as the proposed data split from Section 5.2.3, the first two groups of individuals were combined to form one dataset for training purposes. Opting for a single trainer is a feasible option here since model selection only involves one task (selecting optimal parameters). Thus without the need for validation, the first two groups of individuals were combined to form a larger training base. Consequently the third group of individuals were designated for testing purposes and left untouched until the final application process. This exact same splitting scheme was used across both models.

Note: 80% of the data was used for training - 226 out of 282 subjects. The remaining 20% was used for testing - 56 out of 282 subjects.

5.4.2 Data Preprocessing

Preprocessing and normalization of the HOVON dataset was carried out according to the original protocol, all of which can be seen in Section 5.2.1. The resulting signal intensities were then assumed at the gene level and used for analysis. In particular their logged values (expression) along with the standardized forms were used to construct two separate datasets corresponding to the proposed models. Here the standardization of each feature was done across the training samples.

The objective of proposing multiple datasets (and models) is an attempt to maximize signal processing. Since the logged values and their standardized form can each highlight important genes based on large physical differences and

[31]The RBF kernel along with the vectoral representations of each subject confirms mean-based classification.

congruent collective behavior, the pooled analysis between both forms of data can potentially enhance classification since added differential features are modeled. Thus backed up by added predictive power, the dual system was implemented.

5.4.3 Filtering Procedure

Filtering on both datasets was used to minimize the inclusion of genes associated with low predictive power. In the event of mean-based classification, such features should exhibit similar expression values across both response groups. Hence the proposed filter was designed to remove genes with small differences between their group-wise mean values. Assuming the same set of training subjects from Section 5.2.3, the filter was set up as follows:

- The mean value of each gene corresponding to the response group was calculated (within the training subjects);
- The previous step was repeated for the non-response group;
- The absolute difference between the means were calculated for each gene;
- The absolute differences were ranked and the subset exhibiting the smallest (or largest) difference were removed (or selected).

The filtered subset was purposely associated with ambiguity since it is recognized as an additional parameter in the context of model selection. Therefore the exact number of removed (or selected) genes will remain unknown until the modeling step. However due to computational restraints, the search space pertaining to such removal (or selection) would always be defaulted to the top 1000 genes that exhibit the largest absolute differences. These details are presented in the following section.

5.4.4 Model Selection

The model selection process pertaining to both setups selected the 'best' fit for the training data through parameter optimization. With the designation of the RBF kernel as the inner product and the list of input genes obtained from Section 5.4.3, the cost[32] (C) and cutoff[33] (T) were the only parameters tuned for performance sake. The gamma

[32]Cost parameter used in reference to the SVM setup[20].

[33]Which of the top 1000 filtered genes to use.

parameter (γ) used to adjust kernel depth perception was set to the default[34] due to an augmented search space upon its inclusion. Since the procedure was identical across both models, it would be presented here in a generic fashion by assuming either dataset as the application base.

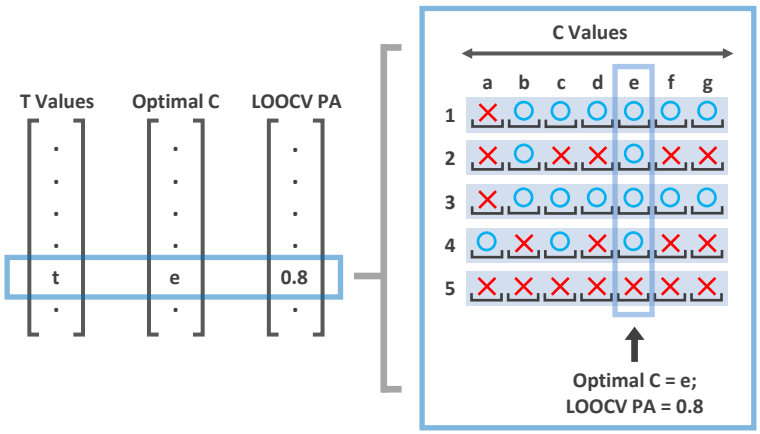


Figure 16: Intuition Behind Mean Modeling

For each instance of T (shown in figure as t), LOOCV was used to generate an objective fit for the training data and consequently select the associated optimal C (shown in figure as e). Since the training dataset was comprised of 226 subjects (shown in figure as 5 subjects labeled $\{1,2,3,4,5\}$), each iteration used a partition of 225 subjects for modeling. The proposed C values (0.1 to 25 in increments on 0.1; 250 in total; shown in figure as $\{a,b,c,d,e,f,g\}$) were then used to construct 250 separate models for each left-out case. Upon completion, the C values where the prediction matched/disagreed with the truth were noted (shown in figure as circles and crosses). Once this process was completed for all 226 subjects, the optimal C was selected by defaulting to a value yielding the largest concordance between prediction and truth (across all 226 subjects). The corresponding proportion (shown in figure as 0.8) is also called the LOOCV prediction accuracy.

To evaluate model performance under various combinations of C and T, selected output using leave one out cross validation[21] (LOOCV) defined the selection criterion across the search space pertaining to both parameters. Here C and T were varied as follows:

- C ranged from 0.1 to 25 in increments of 0.1;
- T ranged from 5 to 1000 in increments of 5.

Note $T = 2$ implies that the top 10 mean difference genes will be used while $T = 20$ implies that the top 100 genes will be used, etc...

[34] Assuming there are p features, then $1/p$ represents the defaulted gamma parameter[20].

After applying LOOCV to each parameter combination, a subsequent duel search was used to locate the ideal pairing. First, the optimal value of C within each instance of T was designated by pinpointing a value that maximized the LOOCV prediction accuracy. Ties here were broken by defaulting to the smallest value so that a C corresponding to the most generalized model was selected[20]. Upon completing this initial step, T was then optimized. Unlike the previous selection process where a lone criterion was used, a combination between LOOCV prediction accuracy, training accuracy, C value, and the number of predicted responses jointly structured this final search. Here multiple criterion ensured objectivity and optimality in the final model. For example by choosing a T that referenced similar LOOCV and training prediction accuracies (and consequently a low C value), objectivity can be achieved by avoiding issues related to overfitting. Furthermore by also favoring a higher number of predicted responses, optimality could also be met since additional true positives should follow in the final application. Note that while this later point will also increase the number of false positives, this problem was partially negated since the final mean predictions will require confirmation across the two models. Thus the implemented strategy falls in line with the goals from Section 5.3: Within each signature, maximize the number of responses; Then use both signatures to confirm the predictions. Figure 16 effectively summarizes the mean selection process.

5.4.5 Individual Model Prediction Result

For the final mean model corresponding to the log intensities ratios, C and T were optimized to 1.9 and 8 (top 40 genes) respectively. The LOOCV prediction accuracy on the training data was 0.7478, which included 26 predicted responses (12 were false positives). The selection of this particular parameter combination was due to the large C value that other models exhibited. Thus despite a larger number of predicted responses, they were avoided due to the risk of overfitting. In terms of the application to the test data, the chosen parameter combination yielded 11 responders (4 false positives) and 45 non-responders (10 false negatives).

For the final mean model corresponding to the standardized intensity ratios, C and T were optimized to 1.7 and 55 (top 275 genes) respectively. The LOOCV prediction accuracy on the training data was 0.7611, which included 27 predicted responses (11 were false positives). The selection of this particular parameter combination was again due to objectivity considerations. In terms of the application to the test data, the chosen parameter combination yielded 12 responders (4 false positives) and 44 non-responders (8 false negatives).

Note that the preprocessing of both test datasets were carried out in the same fashion as their respective training datasets as specified in Section 5.4.2.

5.4.6 Final Mean Prediction Results

The final prediction corresponding to the mean signature of SCRIP was comprised of the intersection between the aforementioned results. Here the intersection was carried out as follows:

- For any given subject, if either (or both) model(s) predicted a non-response, then they were ultimately labeled as a non-responder;
- For any given subject, if both models predicted a response, then they were ultimately labeled as a responder.

Since the goal within each individual signature was to capture as many responders as possible, the proposed strategy falls in line with the overall objectives. Consequently a total of 9 responders (3 false positives) and 47 non-responders (11 false negatives) were obtained.

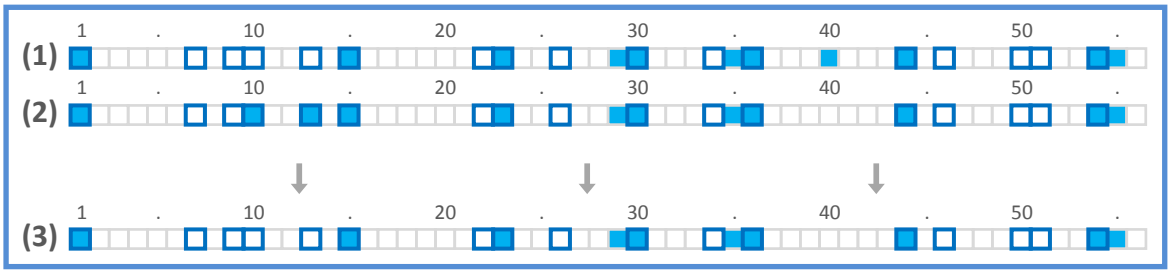


Figure 17: Mean Signature Prediction Result
The individual bars represent the test-subject predictions from the mean models ((**1**) and (**2**)) and signature of SCRIP (**3**). Here (**1**) and (**2**) represents the models derived from the log and standardized ratios while (**3**) represents the union of their results. Within each prediction bar, the boxes highlighted with boarders represent the true responders while the ones shaded in light blue correspond to the predicted responders (the numbers on top of each bar index the subjects within the test dataset). The final mean signature therefore predicted a total of 9 responders (3 false positives) and 47 non-responders (11 false negatives); an accuracy of 0.75.

5.5 Correlation Signature Application Process

The correlation signature of SCRIP represents a novelty proposed in this thesis. Based entirely on the theory from Chapter 3, it attempts to classify graphical representations of each subject through the help of a correlation-based metric. Thus from an approach point of view, it completely differs with respect to its mean-based counterpart.

Nevertheless while the inherent differences between both signatures give rise to their uniqueness, the training strategies for the most part were shared between both applications. For example in order to maximize the number of predicted responses, both signatures use an aggregated input across multiple models to define their final predictions. In the correlation case these models will end up corresponding to the pathways or overall graphs (OG) used within each formulation. Since numerous pathways exist, numerous correlation-based models were consequently trained under this setup.

By construction the training of these models were carried out on the same standardized dataset. In light of this universal prediction base however, different models still exhibited individual characteristics since each OG pinpoints a unique set of co-expression relationships in which the classification is predicated on. Thus the large number of correlation-based models should not be interpreted as a drawback, i.e. the intrinsic value of each model was somehow diffused or mitigated. Instead it should be taken as a thorough attempt to explore all potential co-expression relationships that are capable of differentiating between response groups. Nevertheless getting back to the application process, upon training all models, a majority voting scheme[22] is then used to formalize the final correlation signature of SCRIP.

The following section outlines the proposed application process. This includes the OG selection, data split, data preprocessing, feature filtering, initial training, and result combination.

5.5.1 Selection Of Overall Graphs

Pathways from four publicly available databases were used to curate a list of OGs subsequently analyzed through SCRIP. The databases included Reactome[23], KEGG[24], BioCarta[25], and NCI[26]. In total, 765 pathways were obtained to form a comprehensive list of these gene-wise interactions.

Table 5: Pathway Summarization From `Graphite`

Retrieval Database	Number of Pathways	Mean (Median) Nodes	Mean (Median) Edges
KEGG	130	71.86 (54)	211.12 (75)
Reactome	465	33.22 (14)	780.64 (33)
BioCarta	94	15.18 (14)	36.88 (28)
NCI	76	76.79 (48)	165.18 (81)

Table summarizes the pathways after conversion to networks. The average number of edges and nodes are given according to the retrieval database.

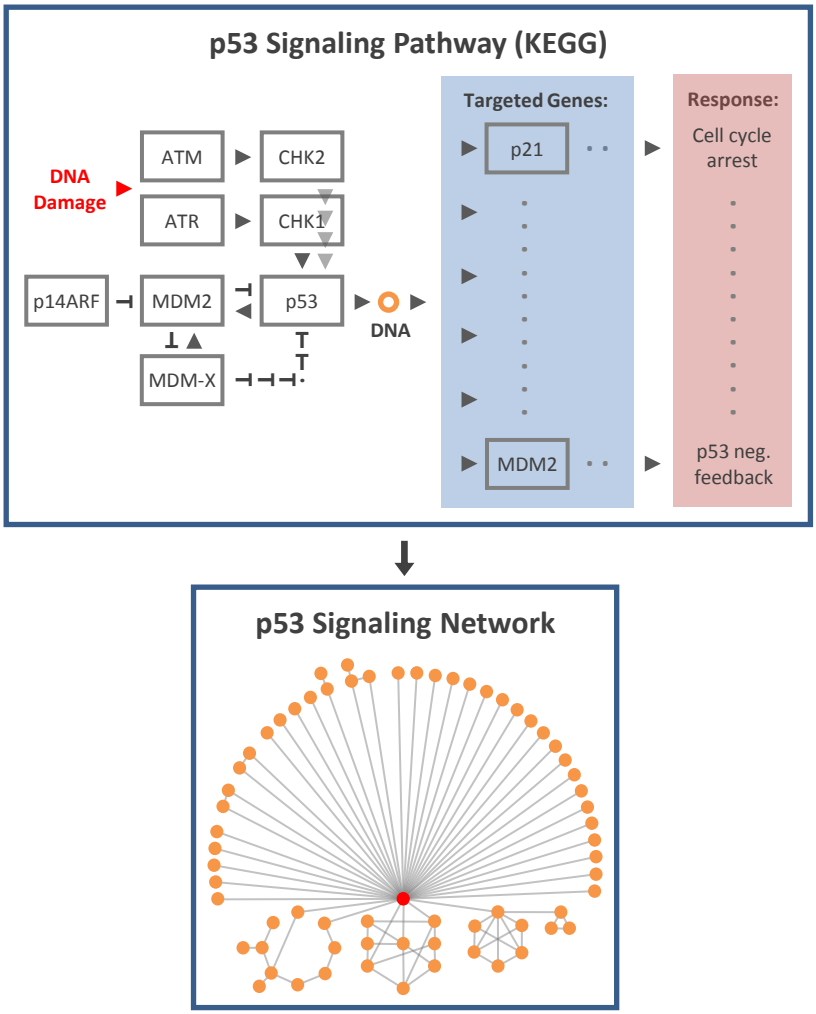


Figure 18: Overall Graph Preprocessing

Flowchart represents the conversion process between biological pathways (top panel) and networks (bottom panel). Initial pathways (top panel) were downloaded from public repositories including KEGG, BioCarta, Reactome, and NCI. Based on various assumptions, the `graphite` package was then used to convert the topological information into a graph based network readily available for use (bottom panel). Additional processing was required to ease computation nonetheless.

To convert these pathways into OGs resembling gene-gene networks (or labeled graphs), the '`graphite`[27]' (**GRAPH I**nteraction from pathway **T**opological **E**nvironment) Bioconductor package was used. This decision was based on its intuitive working assumptions and efficient interaction it provided with respect to SCRIP. For example from the assumption point of view, `graphite` interprets pathway information using biologically-driven rules so that the reconstruction of the corresponding networks will account for protein complexes, gene families, and chemical compounds[27]. As a result the final OGs not only reinterpret pathways for SCRIP, but also retains many nuances from their original form. Thus in light of the straightforward pathway manipulation, the use of `graphite` in conjunc-

tion to SCRIP becomes an obvious choice.

Application wise, all 765 pathways were processed through `graphite` to arrive at their OG representations. Additional details on these pathways can be seen in Table 5.

5.5.2 Data Split

In the same fashion as the proposed data split from Section 5.2.3, the first two groups of individuals were designated for training purposes. Here group one was reserved for parameter optimization and model selection (pertaining to each pathway) while group two then adjusted the majority voting scheme (used to combine the individual fits across all pathways). Thus unlike the mean signature where training was restricted to a single dataset, an additional set of subjects were reserved for validation purposes. Consequently the first two groups of subjects formed the training and validation datasets while the third group was again held out for testing purposes. Again these subjects were left untouched until the final application process.

Note: 60% of the data was used for training - 169 out of 282 subjects. 20% was used for validation - 57 out of 282 subjects. The remaining 20% was used for testing - 56 out of 282 subjects.

5.5.3 Data Preprocessing

Preprocessing and normalization of the HOVON et al. dataset was carried out according to the original protocol, all of which are previously outlined in Section 5.2.1. The resulting signal intensities were then assumed at the gene level and used for analysis. Since the correlation-based models used standardized values as their defaulted input, only one dataset matching this specification was created. Here the standardization was implemented across the training samples; in the same fashion as the dataset from the second mean-based model.

Similar to the mean signature, multiple correlation-based models were trained in anticipation of a concurrent increase in signal level. The justification behind such thought process was as follows: First, since each pathway designates a set of important co-expression relationships, classification will therefore be based on a larger pool of predictive signals given the increase in pathway count; And second, by virtue of having more pathways, informative/relevant co-expression relationships repeatedly designated across multiple pathways will be highlighted and factored into the computation accordingly. Thus for both reasons, training multiple models will most definitely increase the predictive

power corresponding to this particular signature - a goal from the very start.

5.5.4 Filtering Procedure

The filtering pertaining to each correlation-based model was entirely specified by its pathway. Here the important co-expression relationships were designated and subsequently used for classification. Thus unlike its mean-based counterpart, additional training wasn't required.

5.5.5 Initial Training

Initial training of the correlation-based models included parameter optimization followed by pathway selection. In reference to the first step, optimization only pertained to the cost variable (C) since it represented the only free parameter under this setup. Once the procedure was completed, pathway selection then removed (or selected) non-predictive (or predictive) models from the input list of 765 OGs using results from the previous optimization. Here both tasks were carried out on the training data alone.

For all pathways, model performance under different values of C were evaluated using leave one out cross validation (LOOCV). Output from this procedure was then used to define a selection criterion that pinpointed the 'best' fitting model (and therefore an optimal C). Since the implemented search was consistent across all pathways, it would be presented here in a generic fashion. Thus assuming any pathway P and the proposed training data, LOOCV was carried out as follows:

- C was varied between $1,000$ and $2,000,000$;
- For a given individual, regions of C where model prediction matched/disagreed with the true response label were computed using a bisection algorithm[28];
- Across all subjects, the first two steps were repeated;
- Optimal C (of pathway P) was then designated as the value exhibiting the most concordance between prediction and truth across all subjects. This proportion was then referred to as the LOOCV prediction accuracy.

Note that ties in the final step were broken by favoring the smallest C value so that the most generalized classifier could be guaranteed. Once the procedure was repeated across all 765 input pathways, parameter optimization concluded.

With each pathway optimized for performance, model selection was subsequently used to filter away non-predictive inputs/pathways. As a reminder they correspond to non-effective OGs or essentially priors that fail to designate any informative gene pairs[35]. Since the initial input list of 765 pathways would inevitably contain a selection of these targets, their removal becomes a necessity in order to control the noise level associated with the final correlation signature. Here the removal was based on the concept of 'beating the proportion'.

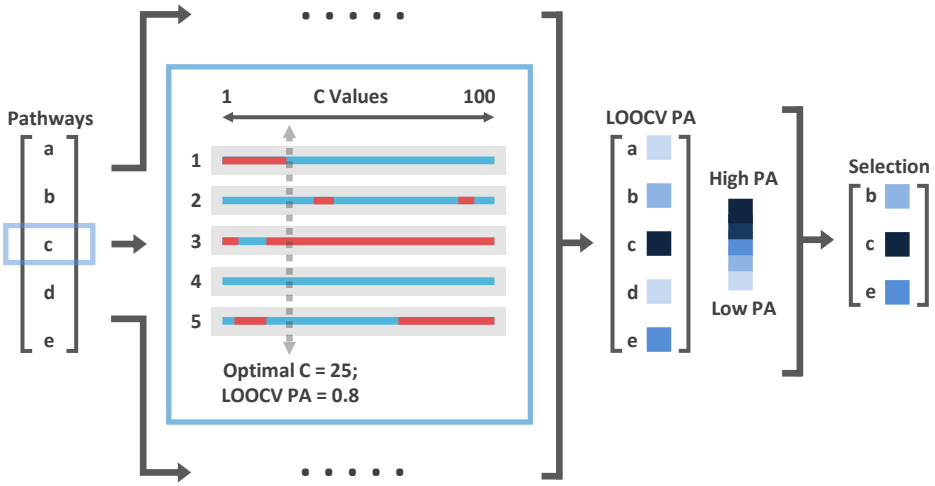


Figure 19: Intuition Behind Initial Correlation Modeling

In the depicted example, assume that a total of 5 pathways $\{a,b,c,d,e\}$, 5 individuals $\{1,2,3,4,5\}$, and C values between 1 and 100 were considered. Starting at the very left panel, the initial training process was comprised of pathway filtering and optimization. In particular each given pathway was optimized for the best 'C' value using LOOCV. (Middle left panel) Specifically for a given individual, a prediction model built on the other four subjects was used to evaluate the predictive ability under various inputs of C. Thus the regions where model prediction matched the true respond label for that given individual was computed. In the figure, regions of C which produced matching predictions were shaded in blue while the non-matching regions were shaded in red. Once this procedure was completed for all subjects, the optimal C was then designated as the smallest value exhibiting the most concordance between prediction and truth across all subjects. This proportion was then referred to as the LOOCV prediction accuracy. (Middle right panel) Once this was carried out across all pathways, parameter optimization concluded. (Right panel) By subsequently using the output from the LOOCV procedure - which would include the prediction accuracy, number of predicted responders, etc... - non-predictive pathways were then filtered out.

In simple terms, 'beating the proportion' refers to a prediction accuracy that each optimized model needed to achieve in order to qualify for selection. Here the proportion was set at a value reflecting blind prediction since pathways lower than this cutoff would be deemed non-predictive and subsequently removed (since they can even beat

[35] An in depth discussion of non-informative gene pairs can be seen in Section X.

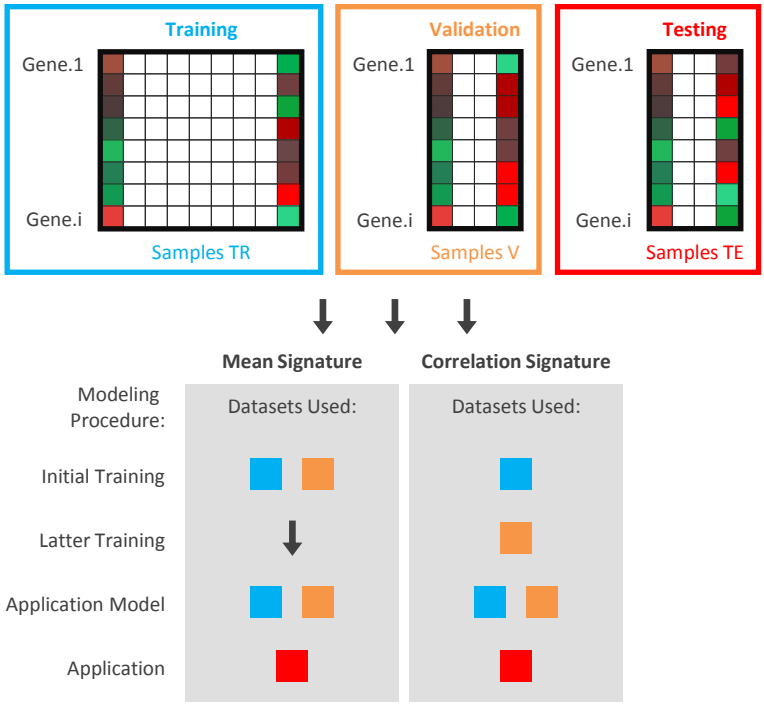


Figure 20: Datasets Used In Application Process

For both signatures of SCRIP, the expression dataset was initially split into training, validation, and testing sections (upper panel). (Lower left panel) For the mean signature, training only comprised of parameter selection. Therefore the training and validation datasets were combined for this purpose. Afterwards the optimal model was also constructed from this same selection. The test data was left untouched until the application process of the proposed signature. (Lower right panel) For the correlation signature, training was comprised of two tasks including parameter selection and pathway combination. Hence the training and validation datasets was used to accomplish both tasks respectively. The optimal model was then built on the combined selection using results from the individual procedures. Again, the test data was left untouched until the application process.

predict based on nothing). While setting the cutoff to 50% would have made intuitive sense, a more sophisticated route involving the maximum responder/non-responder proportion was instead adopted. This decision was made since blind prediction could entail unanimously labeling everyone as a responder or non-responder. Thus in these cases, the maximum proportion is automatically achieved even without modeling or prior information. For example in the training data where 126 of the 169 subjects were non-responders, a unanimous blind prediction would yield an accuracy of $126/169 = 0.7455$. Consequently any model capturing true differential information should be capable of outperforming this random guess and its corresponding percentage.

With this idea in place, the filtering therefore selected pathways exhibiting a minimum LOOCV prediction accuracy of 0.7455. As a result only 202 pathways made it to the final training step. A selection of them can be seen in Table 6.

Table 6: Initial Training Pathway Output

Pathway	C	Accuracy (Training; Validation)	True -	False -	False +	True +	Retrieval Database
Linoleic acid metabolism	21497	0.858; 0.7692	119	32	7	11	KEGG
Sulfur metabolism	17080	0.7811; 0.7633	123	37	3	6	KEGG
P53 signaling pathway	14167	0.8166; 0.7515	121	37	5	6	KEGG
Bone remodeling	548646	0.8047; 0.7633	124	38	2	5	BioCarta
Ccr3 signaling in eosinophils	21108	0.7811; 0.7692	124	37	2	6	BioCarta
Cd40l signaling pathway	532102	0.7751; 0.7633	124	38	2	5	BioCarta
Cell to cell adhesion signaling	1987	0.8225; 0.7633	123	37	3	6	BioCarta
Cyclin e destruction pathway	1046306	0.7988; 0.7692	123	36	3	7	BioCarta
Cystic fibrosis transmembrane conductance regulator and beta 2 adrenergic receptor pathway	4081	0.8107; 0.7692	124	37	2	6	BioCarta
E2f1 destruction pathway	1046306	0.7988; 0.7692	123	36	3	7	BioCarta
Il 2 signaling pathway	109375	0.8462; 0.7751	121	33	5	10	BioCarta
Influence of ras and rho proteins on g1 to s transition	22659	0.8343; 0.7633	120	34	6	9	BioCarta
Mechanism of protein import into the nucleus	4682	0.8047; 0.7692	124	37	2	6	BioCarta
Overview of telomerase rna component gene hterc transcriptional regulation	101413	0.7751; 0.7574	122	37	4	6	BioCarta
Repression of pain sensation by the transcriptional regulator dream	80204	0.8402; 0.7692	120	33	6	10	BioCarta
Role of ß-arrestins in the activation and targeting of map kinases	130542	0.8521; 0.7574	117	32	9	11	BioCarta
Sodd/tnfr1 signaling pathway	12212	0.7811; 0.7692	124	37	2	6	BioCarta
ß-arrestins in gpcr desensitization	439909	0.7811; 0.7751	123	35	3	8	BioCarta
Alternative NF-kappaB pathway	238159	0.8462; 0.7633	117	31	9	12	NCI
C-MYB transcription factor network	9140	0.7988; 0.7574	121	36	5	7	NCI
Calcium signaling in the CD4+ TCR pathway	16532	0.7929; 0.7751	124	36	2	7	NCI
Glucocorticoid receptor regulatory network	4606	0.8047; 0.7633	123	37	3	6	NCI
IL12 signaling mediated by STAT4	519474	0.7988; 0.7633	121	35	5	8	NCI
JNK signaling in the CD4+ TCR pathway	14400	0.8521; 0.7574	121	36	5	7	NCI
Ras signaling in the CD4+ TCR pathway	3940	0.787; 0.7692	125	38	1	5	NCI
S1P2 pathway	10192	0.787; 0.7633	123	37	3	6	NCI
Activation of PKB	1084438	0.8047; 0.7515	118	34	8	9	Reactome
Activation of caspases through apoptosome-mediated cleavage	74203	0.7811; 0.7574	121	36	5	7	Reactome
Apoptotic factor-mediated response	8640	0.7929; 0.7515	120	36	6	7	Reactome
Assembly of the RAD50-MRE11-NBS1 complex at DNA double-strand breaks	111431	0.8047; 0.7515	120	36	6	7	Reactome
Beta oxidation of decanoyl-CoA to octanoyl-CoA-CoA	128078	0.8047; 0.7574	120	35	6	8	Reactome
Citric acid cycle (TCA cycle)	7244	0.8225; 0.7515	120	36	6	7	Reactome
Collagen adhesion via alpha 2 beta 1 glycoprotein	1442352	0.7692; 0.7515	122	38	4	5	Reactome
Cytochrome c-mediated apoptotic response	74203	0.7811; 0.7574	121	36	5	7	Reactome
E2F-enabled inhibition of pre-replication complex formation	19916	0.8047; 0.7515	122	38	4	5	Reactome

Table 6 (Continued): Initial Training Pathway Output

Formation of apoptosome	1689323	0.787; 0.7692	119	32	7	11	Reactome
G2 Phase	1522522	0.787; 0.7574	117	32	9	11	Reactome
G2/M DNA damage checkpoint	1370483	0.7751; 0.7633	124	38	2	5	Reactome
Interleukin-1 processing	253770	0.8166; 0.7574	122	37	4	6	Reactome
Leading Strand Synthesis	11488	0.8047; 0.7574	123	38	3	5	Reactome
MAP kinase activation in TLR cascade	8226	0.8284; 0.7633	121	35	5	8	Reactome
MRN complex relocalizes to nuclear foci	111431	0.8047; 0.7515	120	36	6	7	Reactome
Metabolism of amino acids and derivatives	14446	0.8462; 0.7692	122	35	4	8	Reactome
Metabolism of vitamins and cofactors	6950	0.8107; 0.7515	120	36	6	7	Reactome
Neuroransmitter Receptor Binding And Downstream Transmission In The Postsynaptic Cell	11987	0.8107; 0.7515	122	38	4	5	Reactome
Polymerase switching	11488	0.8047; 0.7574	123	38	3	5	Reactome
Prostacyclin signalling through prostacyclin receptor	832035	0.7811; 0.7633	123	37	3	6	Reactome
Recruitment of NuMA to mitotic centrosomes	7526	0.8166; 0.7692	125	38	1	5	Reactome
Regulation of Lipid Metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	10534	0.7929; 0.7751	124	36	2	7	Reactome
Signal amplification	16410	0.8107; 0.7574	122	37	4	6	Reactome
TRAF6 Mediated Induction of proinflammatory cytokines	9313	0.8047; 0.7692	124	37	2	6	Reactome
Thromboxane signalling through TP receptor	79744	0.8225; 0.7515	122	38	4	5	Reactome
Mitochondrial fatty acid beta-oxidation of unsaturated fatty acids	355949	0.858; 0.7692	117	30	9	13	Reactome

Continued: A selection of the filtered pathways deemed informative from the initial training process (only using the training data) are depicted in the given table. These pathways would subsequently take part in the latter training process where the cutoff for the majority voting scheme is determined. Here the responders are coded as '+' while non-responders as '-'. The validation accuracy refers to the LOOCV prediction accuracy. Note that a total of 202 pathways were deemed informative (produced at least one true responder).

5.5.6 Final Training (Majority Voting)

With the filtered pathways from the previous optimization, final training then formalized the majority voting scheme used in conjunction to each individual model. Here the validation data was used to objectively define the cutoff (T) referenced to in the voting. The computation was carried out as follows:

- Subjects from the validation data (57 total subjects) were preprocessed in the same fashion as subjects from the training data;
- The optimized models from Section 5.5.5 were subsequently applied to the validation data. Here the prediction was carried out on an individual basis for each pathway. Thus 202 separate sets of predictions (for the 57 validation subjects) were obtained;
- Assuming that a response was given a value of 1 and a non-response 0, the summation of these values across all 202 predictions sets was repeated on each validation subject. The resulting summations then formed the 'majority vote vector[36]';
- Based on the results from the majority vote vector, T was set to 10. Therefore a minimum of 10 individual response predictions (from the 202 pathways) were required for any subject to qualify as a responder in the final correlation signature. Ones that fell short of this cutoff were subsequently defaulted as non-responders.

Note that T was selected in an attempt to maximize the number of predicted responses. Here it was set to 10 in order to capture the greater majority of responders. On the other hand it wasn't lowered any further due to the unreasonable trade off with false positives upon its implementation.

With the full specification of the majority voting process finished, the correlation signature was finally ready for application.

5.5.7 Final Prediction Process

Using the optimized C from Section 5.5.5, the filtered pathways were re-fitted to the combined dataset between training ($n = 169$) and validation ($n = 57$) subjects. Thus a total sample size of 226 was used to train the final set of correlation-based models. Upon completion, they were applied to the test subjects ($n = 56$) and the resulting predictions were then combined according to the majority voting scheme derived in Section 5.5.6 ($T = 10$).

---

[36]Each element in the majority vote vector (1x57) ranges between 0 and 202.

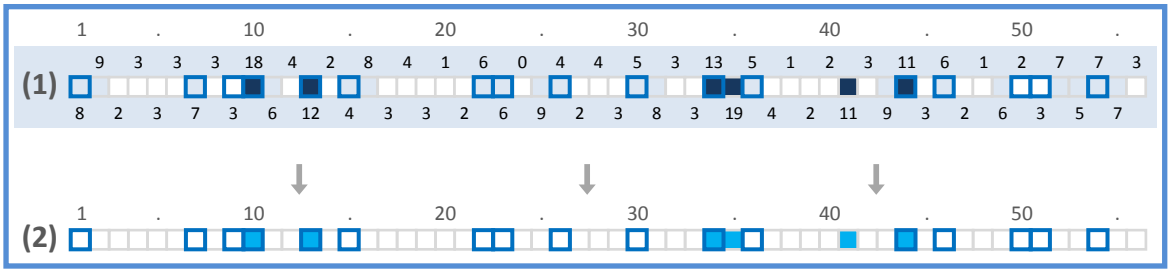


Figure 21: Correlation Signature Prediction Result

The individual bars represent the test-subject predictions from the correlation signature of SCRIP **(2)**. Here **(1)** depicts the vote count (the smaller numbers closer to the bar) corresponding to each individual (the number above each bar index the subjects within the test dataset) and is shaded based on the number of votes each subject received. Subjects that received the most votes (10+) were shaded in a darker hue vs. subjects that received fewer (5 to 9) to an inconsequential total (0 to 4). Nevertheless only the ones that amassed more than 10 votes were labeled as responders in the final predictions ((**2**) - also shaded in blue). Consequently the correlation signature predicted a total of 6 responders (2 false positives) and 50 non-responders (13 false negatives); an accuracy of 0.72.

At the end, 6 responders (2 false positives) and 50 non-responders (13 false negatives) were obtained.

5.6 Final Results

Using the individual outputs from the mean and correlation signatures, the final predictions were subsequently obtained by taking the union of their results. Here the process was carried out as follows:

- SCRIP will predict a non-responder if and only if both signatures labeled the subject as a non-responder;
- SCRIP will predict a responder otherwise.

Using these rules a total of 14 subjects were predicted as responders (4 false positive) and 42 as non-responders (7 false negatives). The resulting accuracy was therefore 0.8036. This effectively concludes the application part of the thesis.

References

[1] R. Kyle, *et al.*, *Mayo Clinic proceedings. Mayo Clinic* (1975), vol. 50, p. 29.

[2] A. Broyl, *et al.*, *Blood* **116**, 2543 (2010).

[3] S. Amin, *et al.*, Gene expression profile alone is inadequate in predicting complete responses in multiple myeloma (2010).

[4] W. Bensinger, *Journal of Clinical Oncology* **26**, 480 (2008).

[5] J. Harousseau, *et al.*, *haematologica* **91**, 1498 (2006).

[6] R. Edgar, M. Domrachev, A. Lash, *Nucleic acids research* **30**, 207 (2002).

[7] J. Wu, R. Irizarry, J. Macdonald, J. Gentry, *R package version* **2100** (2005).

[8] T. Downey, *Methods in enzymology* **411**, 256 (2006).

[9] B. Bolstad, affyplm: Methods for fitting probe level models to affy data (2004).

[10] R. Gentleman, *et al.*, *Genome biology* **5**, R80 (2004).

[11] N. Heintzman, *et al.*, *Nature* **459**, 108 (2009).

[12] R. Team, *et al.*, *R Foundation Statistical Computing* (2008).

[13] X. Liu, J. Wu, Z. Zhou, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **39**, 539 (2009).

[14] S. Ertekin, J. Huang, L. Bottou, L. Giles, *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (ACM, 2007), pp. 127–136.

[15] S. Kotsiantis, I. Zaharakis, P. Pintelas, *Frontiers in Artificial Intelligence and Applications* **160**, 3 (2007).

[16] Z. Zhou, X. Liu, *Knowledge and Data Engineering, IEEE Transactions on* **18**, 63 (2006).

[17] R. Prati, G. Batista, M. Monard, *MICAI 2004: Advances in Artificial Intelligence* pp. 312–321 (2004).

[18] A. Brazma, J. Vilo, *et al.*, *FEBS letters* **480**, 17 (2000).

[19] I. Dhillon, Y. Guan, B. Kulis, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM, 2004), pp. 551–556.

[20] C. Hsu, C. Chang, C. Lin, *et al.*, A practical guide to support vector classification (2003).

[21] J. Shao, *Journal of the American Statistical Association* **88**, 486 (1993).

[22] L. Penrose, *Journal of the Royal Statistical Society* **109**, 53 (1946).

[23] G. Joshi-Tope, *et al.*, *Nucleic acids research* **33**, D428 (2005).

[24] H. Ogata, *et al.*, *Nucleic acids research* **27**, 29 (1999).

[25] D. Nishimura, *Biotech Software & Internet Report: The Computer Software Journal for Scient* **2**, 117 (2001).

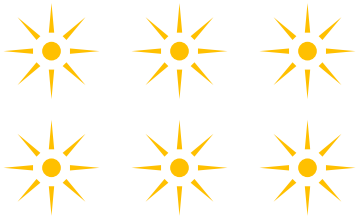
[26] J. Voigt, B. Bienfait, S. Wang, M. Nicklaus, *J. Chem. Inf. Comput. Sci* **41**, 702 (2001).

[27] G. Sales, E. Calura, D. Cavalieri, C. Romualdi, *BMC bioinformatics* **13**, 20 (2012).

[28] Y. Saab, *Computers, IEEE Transactions on* **44**, 903 (1995).

Extension To Copy Number

Chapter 6



This chapter details the extension of '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) to copy number data. Here two separate ideas for the integration process will be provided. The first involves a modification of the current walk kernel while the second proposes new theory in light of the added computation. In both cases, only the background developments are presented. The application was omitted due to the absence of paired copy number data.

6.1 Introduction

As an integrative tool '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) proposes an enhanced classification framework predicated on both mean- and correlation- based signals. By embracing a panoply of approaches, assumptions, and interchangeable procedures, the methodology was purposely designed around flexibility - a unique quality with which many existing tools lack. Consequently due to the implied diversity, it should not come as a surprised that SCRIP commands additional efficacy in comparison to most of its counterparts.

Nevertheless while SCRIP was built to reflect a multitude of approaches, it is still relatively one-sided when it comes to its reliance on the gene expression data. Simply put these profiles will largely determine the success of SCRIP since they represented the only source of farmed signal corresponding to either signature. For example from the correlation perspective, the added pathways only guide the computation and therefore contributes minimally to model derivation, while to a greater extreme, the mean signature even restricts inference to these profiles. Thus it is clear that over-dependency, in reference to these signals, is being exercised.

However due to the drawbacks of modeling response using only one genomic input[1] (especially as it pertains to the expression data in this setting), it behooved the need to incorporate an additional source of information. Because in the event that the expression data ends up being insufficient[2], a valid alternative can then be used to backup the existing analysis and avoid blind prediction if otherwise ignored. Consequently the copy number (CN) data was integrated into SCRIP as planned. Some of the additional motivating factors include:

- The CN data can easily adapt to the presence of the expression profiles and pathways;
- The paired combination between CN and GE were by far the most accessible type of genomic dataset;
- The integration fell in line with the discussion from Chapter 1 under which their merits were detailed.

To therefore carry out the CN analysis within the framework of SCRIP, its formulation and theoretical development represents the only challenges behind the integration process. Fortunately due to the flexibility associated with SCRIP, changes of this nature can all be carried out with relative ease despite the added complexity of introducing an additional data type into the mix. Consequently two contrasting methods are proposed here for the integration process. In the first approach, the CN data will be used as a modification tool to enhance the precision of the correlation workflow; while secondly, a new signature will be altogether proposed for the added data type.

The following chapter therefore details the assumptions and setup of both integrative procedures. Here the discussion will only be presented from a theoretical point of view. The application, in reference to the HOVON[3] dataset (from Chapter 5), was omitted due to the missing CN data.

6.2 Assumptions For CN Data

The copy number (CN) data used in reference to both integrative procedures is assumed in a discrete form. Thus if X represents the CN variable corresponding to some gene, then its domain can be described as: $X \in \{0,1,2,3,\ldots\}$, where the values will reference the number of gene copies.

6.3 First Integrative Procedure

The first integrative procedure was designed as a modification tool with respect to the correlation signature of SCRIP. Here the copy number (CN) data was used in conjunction to the expression profiles as a refinement that increased the distinction and accuracy of the inferred transition probabilities. Thus based on the existing copies of the genes involved, they adjusted the correlation signature (from any given pathway) by highlight various relationships accordingly.

With that being said, this initial integrative procedure is strictly based on logic alone. In other words the interactions between CN and correlation status are assumed to follow a premeditated pattern (free of biological guidance) which consequently defines the underlying computation. Assuming that genes $\{g, g_1, \ldots, g_p\}$ are used in reference to a hypothetical correlation signature, the workflow is carried out by adopting the following notation:

- g: Transition gene (transition from):
- $\{g_1, \ldots, g_p\}$: Potential genes to transition (starting from g);
- $\{p_t(g_1|g), \ldots, p_t(g_p|g)\}$: Transition probabilities corresponding to $\{g_1, \ldots, g_p\}$ (starting from g);
- $p_q(g)$: Ending probability corresponding to g;
- $\{c, c_1, \ldots, c_p\}$: Discretized CN data of $\{g, g_1, \ldots, g_p\}$ according to Section 6.2.

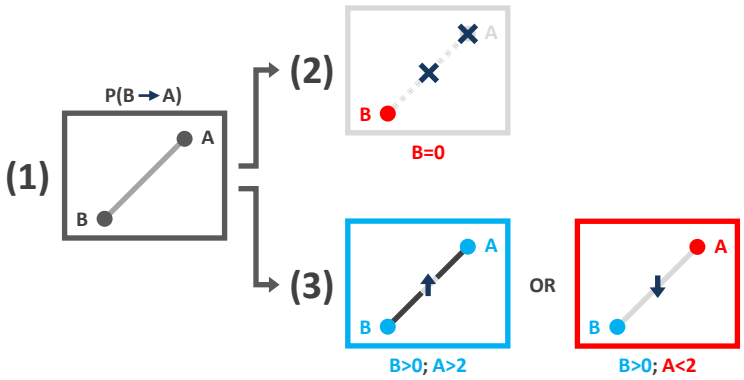


Figure 22: Intuition Behind CN Integration (1)

(1) Assume that inference is restricted to the transition probability from B to A. **(2)** If B is deleted (coded in red), then all links and nodes connecting to B would be removed from the resulting individual graph. In the case that B isn't deleted **(3)**, then the transition probability $P(B \to A)$ would be modified according to the CN status of A (red corresponds to deletion and blue corresponds to amplification). Specifically $P(B \to A)$ would increase if A is amplified and decrease otherwise.

Under this setup, modifications to the ending and transition probabilities are carried out as follows: (changes were reflected on an individual basis within each IG)

- If $c = 0$, then node g and all of its corresponding edges (connecting to $\{g_1, \ldots, g_p\}$) will be removed since no valid targets can be legitimately correlated with a nonexistent gene (g doesn't exist). Hence the ending and transition probabilities will become voided under this scenario;
- If $c > 0$, then the transition probabilities corresponding to $\{g_1, \ldots, g_p\}$ will be modified according to the CN status $\{c_1, \ldots, c_p\}$. Here the modifications are carried out assuming that the CN can directly impact the transitions within a given walk. For example the computation will assist transitions into amplified targets due to an increase of 'transitionalbe' destinations while inhibit the movement in any other scenario. Thus the probabilities will be strengthened or weakened according to the CN status as planned.

To implement this blueprint, the new transition ($p'_t(\ldots|g)$) and ending ($p'_q(g)$) probabilities are constructed as follows:

$$p'_t(g_i|g) = \frac{c_i}{2} p_t(g_i|g) \;\forall\, i \in \{1, \ldots, p\}$$
$$p'_q(g) = 1 - \sum_{i=1}^{p} p'_t(g_i|g) \tag{31}$$

Thus any gene with a CN ratio X in reference to the baseline (2 copies for a particular gene) will get recalibrate according to this proportion. With this result the integration process of the CN data with respect to the correlation signature

concludes.

6.3.1 Implementation

Since the aforementioned modifications are only restricted to the individual graphs (IGs), the new proposal won't invalidate any existing procedure already built into the correlation signature of SCRIP. Thus the corresponding application can still be carried out in the same fashion as detailed in Chapter 3.

6.4 Second Integrative Procedure

The second integrative procedure was specifically designed on behalf of the CN data much like the mean and correlation signatures used in reference to the expression profiles. In particular it was developed as a separate predictor, termed the 'CN signature', that operates as an independent classification tool within the construct of SCRIP. Here its design emphasizes the amplifications and deletions within predetermined gene sets as a mean-based differential signal. Not surprisingly this idea of 'gene sets' will again highlight the concept of 'pathways' and their unique depiction of interconnected gene targets[4].

Undoubtedly the proposed setup of the CN signature will draw comparisons to its correlation counterpart due to their universal use of pathways. And for the most part while these comparisons are valid, the workflows still differed with respect to their computation and logic. For example while both designate an overall graph (OG) as a biological prior, the correlation signature will use it to map informative gene pairs while the CN signature only interpret it from a structural point of view. Consequently the resulting IGs will also differ. Whereas one references the co-expression status, the other designates copy number profiles within an interconnect group of genes. Not surprisingly their kernels will also contrast despite the same generic labels that can be applied to both formulations. Thus in light of all these differences, the CN signature can be recognized as a complete derivative of the existing correlation setup.

The rest of this section will consequently discuss the underlying theory of the new development. This will be presented in two subsections where the first details the merge process and kernel while the second outlines its integration with SCRIP.

6.4.1 Merge Process

The CN signature will use a merge process between the overall graph (OG) and CN profiles to derive separate individual graphs (IGs) with respect to the amplifications and deletions of an individual. This process is designed to remove nodes (and all corresponding edges) that exhibit 'CN statuses in contrast to the given criterion'. For example assuming that the objective is to infer a deletion-based IG, then all genes (and their corresponding links) with observed CN values greater than 2 will be removed accordingly. Similarly the removal will shift to genes with less than 2 copies for the amplification-based counterpart. Thus in both constructs the resulting IGs will portray the designated alteration on top of the imposed OG structure.

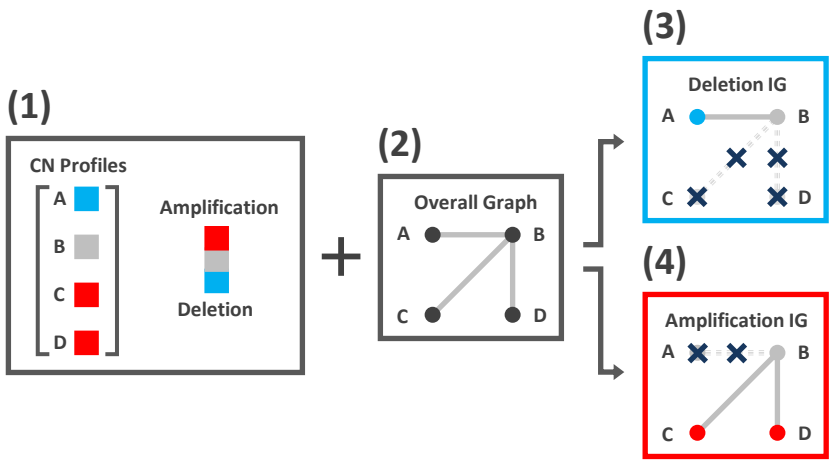


Figure 23: Intuition Behind CN Integration (2)
The second integrative method for copy number data was constructed by defining two separate individual graphs based on the observed genomic profiles **(1)** and the assumed overall graph structure **(2)**. Specifically the amplification based IG **(4)** would display amplifications on top of the imposed structure while the deletion based counterpart **(3)** would display the deletions.

To formalize the described procedure, suppose a hypothetical pathway has the following properties:

- The OG contained j unique nodes $G = \{g_1, \ldots, g_j\}$ (corresponding to j unique genes) and a set of undirected edges E between them;
- $\{c_1, \ldots, c_j\}$ represented a generic copy number profile for $\{g_1, \ldots, g_j\}$.

Under this setup, the two IGs will be described as:

- Alteration IG with G_a nodes and E_a edges where:

$$G_a = \{g_i \in G : c_i \geq 2\}$$

$$E_a = \{e(g_i, g_j) \in E : g_i \in G_a \text{ and } g_j \in G_a\}$$

- Deletion IG with G_d nodes and E_d edges where:

$$G_d = \{g_i \in G : c_i \leq 2\}$$

$$E_d = \{e(g_i, g_j) \in E : g_i \in G_d \text{ and } g_j \in G_d\}$$

Subsequently the described procedure will be repeated across all subjects to form the data summary used in conjunction to the proposed signature.

6.4.2 Kernel Specification

With the specification of the individual graphs (IGs) in place, a valid kernel is subsequently required for their comparative analysis through SVM[5]. Similar to the correlation signature, the implemented kernel also takes advantage of vectoral transformations to define an inner product within the space of graphs[6]. Consequently walks are also used as the metric to evaluate similarity.

Fortunately since the new IGs resemble standard labeled graphs[7] (unlike their previous formulation with probabilities associated to the edges), the implemented kernel can therefore be adopted from a preexisting formulation to avoid many complications with a new derivation. In an attempt to simultaneously maximize efficacy and minimize unnecessary effort, the n^{th} ordered walk function[8] was therefore designated as the inner product in reference to the CN signature. And as its name may already suggest, IGs will be strictly compared through walks of length n. Thus sharing a greater proportion of length n walks will define 'similarity' in this context.

Therefore using the definition of graph structure and proposed theory of product graphs all from Section 3.6, the n^{th} ordered walk function $K_{\text{n.walk}}(\ldots)$ between graphs G_1 and G_2 is described as:

$$
\begin{aligned}
K_{\text{n.walk}}(G_1, G_2) &= \sum_{s \in S(G)} \phi_s(G_1) \cdot \phi_s(G_2) \\
&= \sum_{w_1 \in \mathscr{W}(G_1)} \sum_{w_2 \in \mathscr{W}(G_2)} \lambda_{G_1}(w_1) \cdot \lambda_{G_2}(w_2) \cdot [\text{label } w_1 = \text{label } w_2] \\
&= \sum_{w_1 \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} \sum_{w_2 \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} \lambda_{G_1}(w_1) \cdot \lambda_{G_2}(w_2) \\
&= \sum_{w \in \{\mathscr{W}(G_1) \cap \mathscr{W}(G_2)\}} c_w \cdot \lambda_{G_1}(w) \cdot \lambda_{G_2}(w) \\
&= \sum_{w \in \mathscr{W}(G_1 \times G_2)} \lambda_{G_1 \times G_2}(w) \\
&= \sum_{n=1} \sum_{w \in \mathscr{W}_n(G_1 \times G_2)} \lambda_{G_1 \times G_2}(w) \\
&= \sum_{w \in \mathscr{W}_n(G_1 \times G_2)} 1 \\
&= \mathbf{1}^T \mathbf{A}^n \mathbf{1}
\end{aligned} \tag{32}
$$

, where: $\lambda_G(w) = \begin{cases} 1 & \text{if length of } w \text{ is } n \\ 0 & \text{otherwise} \end{cases}$

$\mathbf{A}$ is the adjacency matrix of $G_1 \times G_2$.

Since two IGs corresponding to the amplification and deletion status are defined for each subject, the final application kernel is therefore constructed as a weighted summation of $K_{\text{n.walk}}(\ldots)$ over both graph sets[9][37]. Assuming that (G_i^a, G_i^d) and (G_j^a, G_j^d) corresponding to the (amplification, deletion) graphs of subject i and j respectively, the final kernel $K^f_{\text{n.walk}}(\ldots)$ is then described as follows:

$$
K^f_{\text{n.walk}}(\text{subject } i, \text{subject } j) = \frac{1}{2} K_{\text{n.walk}}(G_i^a, G_j^a) + \frac{1}{2} K_{\text{n.walk}}(G_i^d, G_j^d) \tag{33}
$$

6.4.3 Integration With SCRIP

The training process of the CN signature will closely resemble its correlation counterpart. Since the new development will also permit multiple input pathways (example can be seen in the application part of the thesis), a majority voting scheme will therefore be required to bridge the individual predictions with the final results. However due to the application specific nature of this step, it will be left up to user interpretation. Hence these details are left out of the discussion.

[37] This was possible since the weighted summation of two legitimate kernels still remains valid.

Nevertheless assuming that the final predictions corresponding to the CN signature can be obtained, its integration with SCRIP will follow a simple confirmation process across the mean and correlation signatures respectively. Assuming that P_{cn}, P_{m}, and P_{cr} corresponds to the prediction results from the CN, mean, and correlation signatures, their combined prediction and ultimately the output of SCRIP will therefore be obtained as follows:

- For any subject if two or more predictions out of P_{cn}, P_{m}, and P_{cr} return a response label, then they will be classified as a responder;
- In any other scenario they will be classified as a non-responder.

References

[1] L. Van't Veer, R. Bernards, *Nature* **452**, 564 (2008).

[2] S. Amin, *et al.*, Gene expression profile alone is inadequate in predicting complete responses in multiple myeloma (2010).

[3] A. Broyl, *et al.*, *Blood* **116**, 2543 (2010).

[4] S. Maddika, *et al.*, *Drug resistance updates* **10**, 13 (2007).

[5] C. Hsu, C. Chang, C. Lin, *et al.*, A practical guide to support vector classification (2003).

[6] A. Smola, R. Kondor, *Learning theory and kernel machines* pp. 144–158 (2003).

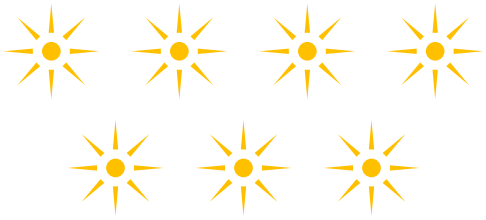
[7] H. Kashima, K. Tsuda, A. Inokuchi, *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-* (2003), vol. 20, p. 321.

[8] S. Vishwanathan, N. Schraudolph, R. Kondor, K. Borgwardt, *The Journal of Machine Learning Research* **99**, 1201 (2010).

[9] K. Parthasarathy, K. Schmidt, *Positive definite kernels, continuous tensor products, and central limit theorems of probability theory* (Springer-Verlag, 1972).

Conclusion

Chapter 7



This final chapter provides a conclusion for this thesis and consequently '**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP). The discussion starts with an interpretation of the application results from Chapter 5 before shifting to future plans in reference to this method.

7.1 Introduction

'**S**VM for **C**omplete **I**ntegrative **R**esponse **P**rediction' (SCRIP) was developed as a novel classification scheme on behalf of this thesis. With a central goal of enhancing the capabilities associated with response-based modeling, its design highlights a series of 'solutions' in response to the shortcomings of existing methods, some of which included: (1) Limitations inherent to the genomic data type[1–4]; (2) Questionable analytical decisions committed by the methodology[5]; and (3) Overly rigid specifications with regards to model training[5].

In response to these problems, SCRIP adopted a comprehensive modeling approach fixated on 'data integration'. And for a variety of analytical reasons, the decision to do so was justified by the presumed payoff when carried through. For example since data integration enables the simultaneous analysis of multiple data types, it will help alleviate some of the insufficiency associated with a single genomic source[4]. Consequently due to the added dimension of such setup, it will also allow the freedom to explore additional approaches towards the problem. Thus similar to a domino effect, the inclusion of data integration naturally infuses the methodology with the needed flexibility and efficacy to tackle a response prediction problem. Hence it represents an ideal setup within the context of SCRIP.

With the finalization of the methodological setup, the application of SCRIP was then conducted on a multiple myeloma[6] (MM) dataset which featured a set of newly diagnosed patients (treated according to a three-drug regimen of VAD/PAD followed by autologous stem cell transplantation[7]). Here these subjects were recruited as part of the HOVON-65/GMMG-HD4[8] trial in an attempt to evaluate the efficacy of bortezomib with respect to drug response and progression-free survival. Since the evaluation process was defined by genomic profiling, the trial consequently collected expression data along with EBMT-derived-response-statuses. Hence it was selected in part of its matching specifications with regards to SCRIP.

In additional to the qualifying practicality of this application, the decision to model MM-based response also reflects the benefits from an analytical point of view. Because without the presence of these advantages, the difficulty associated with this particular application[1] would have warranted an alternative route in light that response data from more favorable cancer types exist - i.e. breast cancer. For example a quick search for breast cancer datasets on public repositories such as Gene Expression Omnibus[9] (GEO) will return up to 1000 response oriented studies all featuring the necessary genomic profiling. And since it has been shown that certain subtypes of breast cancer (i.e. HER2 overexpression) respond favorably under targeted treatment[10–14] (such as Herceptin, NeuVax, and Cetuximab), such applications will seemingly offer more assurance to obtain satisfactory response signatures when modeled according to the designated gene targets. Hence they can be recognized as 'easier' applications in comparison to other

cancer types, i.e. MM, where the same complimentary biological links have yet to be discovered. Ultimately as it will pertain to SCRIP because of its spotlight focus on pathways and gene sets, conducting an analysis without a legitimate target[38] and consequently an absence of confirmed pathways will presumably hinder the otherwise likely success.

Nevertheless despite the uphill battle involved with MM, the application still offered an adequate exchange of upside to balance out its inevitable complications. As a result it justified the decision for this particular cancer type even when other options, potentially more suitable for response prediction, exist (similar to the aforementioned example). With that being said, some of the motivating factors include:

- Large sample size: With a total of 282 subjects, the HOVON dataset was reasonably sized in comparison to other genomic trials. Therefore it offers an unmatched degree of robustness for response prediction purposes.
- Concrete evaluation criteria: While the complications associated with MM supposedly caps prediction accuracy at 70%[1], the impediment was however recognized as an assistance under the construct of this thesis. Specifically the benchmark provides an objective evaluation of SCRIP such that any returned accuracy exceeding 70% will become a simple confirmation of its merit and worth.

Therefore under the presence of these benefits, the MM application was carried out as detailed in Chapter 5.

7.2 Results And Interpretation

The application of SCRIP on the HOVON dataset was carried out in a two-step procedure that separately featured mean- and correlation-based modeling. The training was conducted on an individual basis in order to accommodate methodological flexibility; whether it corresponds to the ease of integrating new signatures or the modifications required to optimize existing ones (i.e. modify the training process to counteract class imbalance). Nevertheless for this particular application, the latter case applied and therefore two sets of predictions (mean and correlation) were trained in reference to the test subjects. Note that while the final results require a combined input across both inference sets, their discussion is presented on an independent basis to allow for a finer degree of interpretation.

[38] In reference to the HER2+ breast cancer subtype, the use of pathways involving HER2 could potentially benefit prediction due to its confirmed pertinence.

7.2.1 Mean Signature Focus

Corresponding to the mean signature of SCRIP, a total of 9 responders (3 false positives) and 47 non-responders (11 false negatives) were obtained in reference to the test dataset (which featured 17 and 39 true responders and non-responders respectively). Surprisingly the workflow yielded a prediction accuracy 0.75 (eclipsing the aforementioned benchmark) despite implementing a limited mean-based approach. This improvement, in light that the signature is identical to any SVM model, is most likely the result of the unique training process specifically adopted for the HOVON dataset (the class imbalance issues as mentioned back in Section 5.3). With that being said while randomness could have also triggered this improvement, it is however unlikely due to the marginal likelihood associated with such event[39]. Thus in this context, the merit of SCRIP receives an initial vote of approval.

From a biological perspective, the genes that contributed to the mean signature were also examined to identify molecular features that can potentially underly drug sensitivity. Here a total of 80 genes corresponding to the largest differentially expressed features were directly corralled from the filtered lists of both mean-based models (logged expression values and their standardized forms). Their subsequent analysis yielded some intriguing findings despite the overwhelming randomness associated with these features. Most notably:

- Amongst the entire list, CCDC104, KDELR3, ARSB, BNIP3, XIST, and UCHL1 exhibited the largest differential expression in the favor of responders;
- Amongst the entire list, HERC6, ZNF202, IRF7, NLGN4X, DDAH1, and KLHL14 exhibited the largest different expression in the favor of non-responders;
- Function analysis of differentially expressed genes in the favor of non-responders (32 total genes) indicated a significant presence of targets involved in the 'response to virus' biological process (6 total targets, 1.5E$-$4 Benjamini adjusted P-value). Similarly the 'endoplasmic reticulum' cellular component was highlighted amongst responders (12 total targets out of 51 genes; 6.9E$-$3 Benjamini adjusted P-value);
- A prominent characteristic in non-responders was the elevated expression of genes controlling 'response to virus'. These included the IRF and CAS families; and could have potentially facilitated drug resistance.

---

[39] A unanimously blind prediction of non-responders would result in a prediction accuracy of 0.69 (39 out of the 56 subjects). Consequently a net of at least 3 correctly classified responders would be required to match the 0.75 obtained through application. With this in mind the probability of observing such event under a random guess could be described as: $\sum_{i=1}^{17} \frac{17 \text{ choose } i}{56 \text{ choose } i} = 0.03$. Thus the possibility that randomness dictated the results could be effectively ruled out.

Note that the partial gene list could be seen in Table 7 and 8.

Table 7: Gene Targets For Mean Model (1)

(A)

Rank	Symbol	Entrez ID	Differential
3	CCDC104	112942	-0.5977
4	KDELR3	11015	-0.5684
5	ARSB	411	-0.5669
6	CTBS	1486	-0.5368
7	SCPEP1	59342	-0.5358
8	CD302	9936	-0.5358
9	GGCX	2677	-0.5316
10	ARSB	411	-0.5312
11	PDIA5	10954	-0.5301
12	BNIP3	664	-0.5286
14	CCNC	892	-0.5072
17	FKBP7	51661	-0.4991
18	MCFD2	90411	-0.4990
19	RP11	25911	-0.4920
21	BRP44L	51660	-0.4864
22	GCSH	2653	-0.4863
23	RECK	8434	-0.4857
24	LY96	23643	-0.4851
28	ALG1	56052	-0.4818
29	HDDC2	51020	-0.4802
31	C6orf89	221477	-0.4783
32	KCTD20	222658	-0.4780
33	TMEM30A	55754	-0.4779
34	CNPY2	10330	-0.4772
35	IQCK	124152	-0.4746
36	DSTN	11034	-0.4739
37	STRAP	11171	-0.4731
38	RAB13	5872	-0.4729
39	ATRN	8455	-0.4727
40	SIL1	64374	-0.4714

(B)

Rank	Symbol	Entrez ID	Differential
13	HERC6	55008	0.5135
16	ZNF202	7753	0.5038
20	IRF7	3665	0.4900
25	IRF9	10379	0.4845
26	NAV2	89797	0.4844
27	MX2	4600	0.4829
30	CXXC1	30827	0.4789

The entries represent the top 40 differentially expressed genes from the mean model (trained on the standardized log expression values). The 'differential' column represents the mean differences between expression values from both response groups, i.e. Mean of non-responders - Mean of responders. Hence **(A)** represents the entries in favor of the responders (negative difference) and **(B)** the ones in favor of the non-responders. Note that these values were calculated from the training data.

7.2.2 Correlation Signature Focus

Corresponding to the correlation signature of SCRIP, a total of 6 responders (2 false positives) and 50 non-responders (13 false negatives) were obtained in reference to the test dataset (which featured 17 and 39 true responders and non-responders respectively). The inflation of false negatives witnessed here was primarily due to the class im-

Table 8: Gene Targets For Mean Model (2)

(A)

Rank	Symbol	Entrez ID	Differential
2	BNIP3	664	-1.6509
3	XIST	7503	-1.4012
6	KDELR3	11015	-1.2637
8	UCHL1	7345	-1.2411
9	TSPAN7	7102	-1.2086
12	KDELR3	11015	-1.1282
13	ERAP2	64167	-1.1078
14	XIST	7503	-1.0881
17	XIST	7503	-1.0363
19	DNAJC12	56521	-1.0214
27	KDELR3	11015	-0.9389
32	BCAT1	586	-0.9106
33	CRIM1	51232	-0.9095
35	TMEM200A	114801	-0.9035
36	CCDC104	112942	-0.9001
37	ICAM4	3386	-0.8900
38	TMEM45A	55076	-0.8823

(B)

Rank	Symbol	Entrez ID	Differential
1	NLGN4X	57502	1.6653
4	DDAH1	23576	1.3283
5	KLHL14	57565	1.2754
7	IFI44L	10964	1.2591
10	IFIT1	3434	1.1937
15	RSAD2	91543	1.0723
16	IFIT3	3437	1.0713
20	IGHG1	3500	0.9893
21	LAPTM5	7805	0.9874
22	HERC6	55008	0.9700
23	MX2	4600	0.9607
24	FRZB	2487	0.9460
25	EPHB1	2047	0.9442
26	SHISA2	387914	0.9414
28	BBOX1	8424	0.9365
29	XAF1	54739	0.9357
31	IFI27	3429	0.9278
34	ANK3	288	0.9060
39	ETS1	2113	0.8795
40	TRIM14	9830	0.8735

The entries represent the top 40 differentially expressed genes from the mean model (trained on the log expression values). Refer to Table 7 for an interpretation.

balance issue. In particular the common cases (in this case the non-responders) are highlighted at the expense of over-predicting their existence. Consequently the application suffered resulting in an accuracy of only 0.72; close to the aforementioned benchmark.

At the bottom line while this performance is clearly undesirable, it also comes as no surprise given the nature of the implemented training. Thus practically speaking, the low prediction accuracy could have been anticipated even before the application finalized. Ironically however the foreseeable complication won't end up compromising the application contrary to what logic may have otherwise suggested. Instead it actually assisted SCRIP by foreshadowing the influx of false negatives. Thus in other words it offers an opportunity to counteract their presence and plan for their impact accordingly. And since this was partially achieved by imposing a filter with respect to both signatures of SCRIP[40], these ramifications were for the most part neutralized leaving the final predictions largely unaffected.

Nevertheless despite the net benefit of such trade off, the fact remained that the correlation signature was clearly

[40]SCRIP takes advantage of a confirmation process between the mean and correlation signatures in order to filter away a subset of the inflated false negatives.

Table 9: Final Training Pathway Output

Pathway	C	Accuracy (Training; Validation)	True -	False -	False +	True +	Retrieval Database
Adipocytokine signaling pathway	32140	0.8894; 0.7655	153	39	14	20	KEGG
Citrate cycle (TCA cycle)	6248	0.7876; 0.7655	166	52	1	7	KEGG
Huntington's disease	23277	0.8230; 0.7655	161	47	6	12	KEGG
Nephrin interactions	3122	0.7743; 0.7655	167	53	0	6	Reactome
Role of ß-arrestins in activation and targeting of kinases	93200	0.8009; 0.7611	162	49	5	10	BioCarta
CD28 dependent Vav1 pathway	1533220	0.9204; 0.7434	142	33	25	26	Reactome
Adipocytokine signaling pathway	32140	0.8894; 0.7655	153	39	14	20	KEGG
Effects of Botulinum toxin	572996	0.885; 0.7522	150	39	17	20	NCI
AKT phosphorylates targets in the cytosol	1519187	0.8274; 0.7522	153	42	14	17	Reactome
Corticosteroids and cardioprotection	193487	0.8628; 0.7478	152	42	15	17	BioCarta

The filtered pathways deemed informative from the initial training process (only using the training data) were refitted to a large training base comprised of both training and validation subjects (before application to the test data). Depicted above are the top candidates based on (cross) validation accuracy and the number of (cross validation) true responders. Here the responders are coded as '+' while non-responders as '-'. Note that a total of 189 pathways were deemed informative (produced at least one true responder).

depreciated in favor of the final application results. Consequently interpretations of the workflow, whether it corresponds to the predictions or biological findings, require additional verification due to the presence of excess noise. With that being said, they include:

- The Adipocytokine signaling, Citrate cycle (TCA cycle), Huntington's disease, and Nephrin interaction pathways exhibited the best cross validation (training) accuracies with ties at 0.7655;
- The CD28 dependent Vav1, Adipocytokine signaling, and Effects of Botulinum toxin pathways predicted the most CV responders at 26, 20, and 20 respectively. Note that there were a total of 59 responders in the training dataset.
- NF-κB related pathways including NF-κB activation/survival and p75NTR signals via NF-κB appeared in the filtered list. This indicated that the activation of NF-κB could contribute to response status in addition to tumor progression[15–17].
- Prominent cancer pathways including AKT signaling, VEGF signaling, Apoptosis, EGFR signaling, and p53 were all featured as OGs in the correlation signature.
- The most commonly highlighted gene pairs from the filtered pathways included internal relationships between the NUP gene family (i.e. NUP135 to NUP210) and their external connections to other gene targets (i.e. AAAS, POM121C, RAE1, RANBP2, and TPR). This suggests that the co-expression patterns of the Nucleoporin genes

can potentially highlight response information with respect to a subset of MM patients.

Some of these pathways are listed in Table 9.

7.2.3 Final Results

Corresponding to the final predictions from SCRIP, a total of 14 responders (4 false positive) and 42 non-responders (7 false negatives) were obtained in reference to the test dataset (which featured 17 and 39 true responders and non-responders respectively). The resulting accuracy of 0.80 confirms the effectiveness of the filter since improvements on top of both individual signatures were witnessed. Thus in the greater scheme of response prediction, these results verify the efficacy and value of SCRIP by surpassing the proposed benchmark with a comfortable margin.

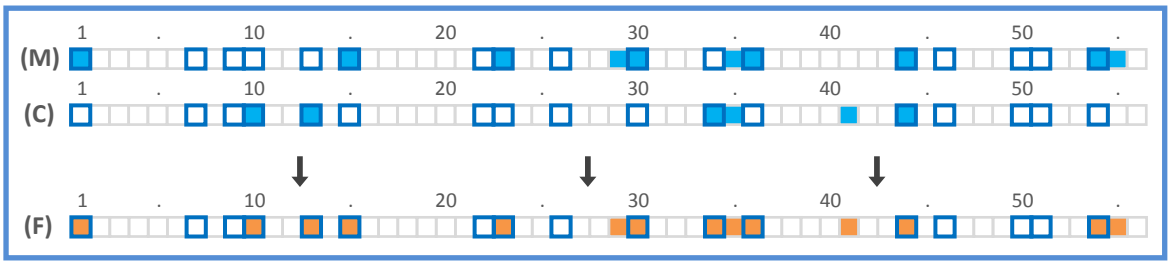


Figure 24: Final Prediction Result

The individual bars represent the test-subject predictions from the **(1)** mean and **(2)** correlation signatures of SCRIP. Here **(3)** visualizes the intersection of their predictions and consequently the final results - 14 total responders (4 false positives) and 42 non-responders (7 false negatives); an accuracy of 0.80. The interpretation follows the same guidelines as in Figure 17.

7.3 Shortcoming And Future Work

In this thesis, it was shown that SCRIP can predict patient response in a MM setting with close to 80% accuracy thus making it a promising candidate for future response-based applications. From a theoretical point of view the success witnessed here is most likely a consequence of the targeted setup implemented in SCRIP: a design predicated on the weaknesses of existing methods. For example by adopting an integrative framework through mean- and correlation-based signals, the proposed methodology was capable of accounting for the insufficiency and rigidness associated with many GEP-based techniques. Consequently this also enables the flexibility to designate an optimal

training strategy on an application specific basis. Thus as demonstrated on the HOVON dataset, SCRIP clearly presents an ideal and realistic framework for response prediction purposes.

Despite these arguments however, SCRIP is far from a finished product in light that the methodology can still use plenty of help. In fact as a relatively conservative technique, most of these improvements will presumably start with the training process and its limited power to yield additional responders as seen in the application. Because even though the methodology optimizes accuracy, it does so at a steep cost involving the sensitivity of the analysis ($10/17 = 0.59$). In particular the responders ended up being predicted at a lower rate in order to accommodate the influx of non-responders in the dataset. While this ensures a greater proportion of accurate predictions (since there are more non-responders), it however contradicts the basic principles of response prediction especially as it pertains to cancer therapeutics; that is to always highlight the potential beneficiaries ahead of all other considerations[18, 19]. Unfortunately this capability was not completely explored in this thesis as it was only brought up in retrospect to the finalized results.

Nevertheless whether or not SCRIP can achieve these lofty goals will remain largely inconsequential as it pertains to this thesis. Because at the bottom line, SCRIP represents more than just another prediction tool. Instead it is meant as an alternative though process that recognizes the value of interchanging multiple data types and approaches in order to restructure a once confined research avenue. And considering the fact that the future will presumably hold a wealth of novelties and genomic sources each with the potential to capture response traits, the initiation to conduct an all inclusive analysis ultimately highlights the key development and lasting contributions hopefully brought forth through this presentation.

References

[1] S. Amin, *et al.*, Gene expression profile alone is inadequate in predicting complete responses in multiple myeloma (2010).

[2] T. Sørlie, *et al.*, *Molecular cancer therapeutics* **5**, 2914 (2006).

[3] B. Ghadimi, *et al.*, *Journal of Clinical Oncology* **23**, 1826 (2005).

[4] L. Van't Veer, R. Bernards, *Nature* **452**, 564 (2008).

[5] D. Ewins, *Modal testing: theory, practice and application*, vol. 2 (Research studies press Baldock, 2000).

[6] F. Zhan, *et al.*, *Blood* **108**, 2020 (2006).

[7] J. Harousseau, *et al.*, *haematologica* **91**, 1498 (2006).

[8] A. Broyl, *et al.*, *Blood* **116**, 2543 (2010).

[9] R. Edgar, M. Domrachev, A. Lash, *Nucleic acids research* **30**, 207 (2002).

[10] C. Vogel, *et al.*, *Journal of Clinical Oncology* **20**, 719 (2002).

[11] J. Baselga, L. Norton, J. Albanell, Y. Kim, J. Mendelsohn, *Cancer research* **58**, 2825 (1998).

[12] B. Squibb, *J Clin Oncol (Meeting Abstracts)* (2008), vol. 26, p. 1009.

[13] K. Pritchard, *et al.*, *New England Journal of Medicine* **354**, 2103 (2006).

[14] M. Scaltriti, *et al.*, *Journal of the National Cancer Institute* **99**, 628 (2007).

[15] T. Hideshima, *et al.*, *Oncogene* **20**, 4519 (2001).

[16] L. YinJun, J. Jie, W. YunGui, *Leukemia research* **29**, 99 (2005).

[17] J. Keats, *et al.*, *Cancer cell* **12**, 131 (2007).

[18] W. Bellamy, *et al.*, *Drugs* **44**, 690 (1992).

[19] T. Hall, *Prediction of Response in Cancer Therapy: Symposium Held Febr. 19-21, 1970, Cascades Conference Center, Williamsburg, Va. Sponsored by Cancer Clinical Investigation Review Committee, National Cancer Institute, National Institutes of Health, Bethesda, Md. and University of Rochester. Editor: Thomas C. Hall* (National Cancer Institute, 1971).